REVISITING AMARI'S α-DIVERGENCE TO IMPROVE CLIPPED SURROGATE PPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Proximal Policy Optimization (PPO) is one of the most popular policy gradient algorithms due to its widespread practicality and scalability. However, its performance is known to be highly sensitive to increasing the clipping threshold and policy network depth. To address this sensitivity, we introduce ALDIPPO, which replaces the clip with an α-divergence penalty on the importance ratio. This makes the penalty stiffer for samples outside the ϵ range, which PPO rejects by switching their gradient off, wasting the rollout budget spent to collect them. We devise several online update rules for α based on on-policy batch statistics, and show that ALDIPPO beats PPO from toy MiniGrid environments to standard MuJoCo tasks, and generalizes better in the Procgen benchmark.

1 INTRODUCTION

Policy-gradient methods update a parametric policy π_θ using data collected by a previous policy $\pi_{\theta_{\text{old}}}$. A key question in these algorithms is *how much can we reuse data collected from* $\pi_{\theta_{\text{old}}}$ *in each update cycle?* With too little reuse, we waste the rollout budget spent to collect the samples. On the other hand, if we reuse them too much, we risk moving the policy too far away from $\pi_{\theta_{\text{old}}}$, which leads to staleness from the collected data. Proximal Policy Optimization (PPO) algorithms (Schulman et al., 2017) address this question by clipping the importance ratio between π_θ and $\pi_{\theta_{\text{old}}}$, nullifying the update gradients past a ratio limit defined by a clip hyperparameter ϵ.

PPO's solution is effective and highly scalable, which has led to its widespread adoption as the go-to on-policy reinforcement learning (RL) algorithm. However, its performance is known to be highly sensitive to the choice of ϵ (Ilyas et al., 2020; Engstrom et al., 2020; Wang et al., 2020) and the depth of the policy network (Xie et al., 2025). Although PPO's clipped surrogate is designed to limit the policy updates, it does not represent a hard constraint on the updates, allowing the policy to drift arbitrarily in the zero-gradient region. This can lead to suboptimal updates and poor performance, especially in deeper networks where the mapping from parameters to policy can be highly nonlinear.

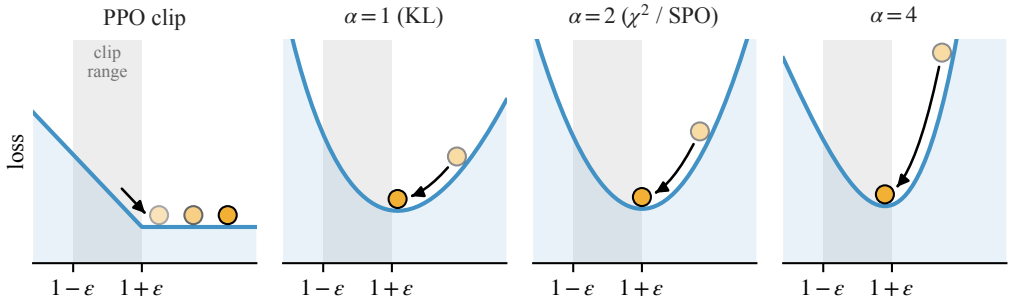

Figure 1: Conceptual visualization of the update geometry of PPO and ALDIPPO for a sample with positive advantage, where $\alpha = 1$ corresponds to $D_{\text{KL}}[\pi_\theta \| \pi_{\theta_{\text{old}}}]$. PPO's clipped surrogate has a zero-gradient region, letting the policy drift arbitrarily past the clip boundary. In contrast, an α-divergence penalty has a non-zero gradient outside of the target, allowing for more effective updates. Crucially, ALDIPPO can adjust the penalty's geometry through the α parameter, which can be tuned online based on the batch statistics.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Prior work has addressed this issue by replacing the clip with a smooth penalty on the importance ratio, such as the quadratic penalty used in Simple Policy Optimization (SPO) (Xie et al., 2025). This penalty has a non-zero gradient everywhere, adding a learning signal that provides control over the policy's movement outside the ϵ range. However, SPO's penalty has a fixed geometry that may not accommodate the changing statistics of the collected data during training.

In this work, we generalize the penalty to the α-divergence family, which allows for a tunable geometry that can be adjusted online based on the batch statistics. This leads to our proposed method, ALDIPPO, which replaces PPO's clip with an α-divergence penalty on the importance ratio, providing better control over the policy's movement through the α parameter. The contrast between PPO's update geometry and a penalty from the α-divergence family is visualized in Figure 1. From this, our contributions are summarized as follows:

1. We show that PPO's clip does not bound the importance ratio when samples in a batch interact through shared parameters, and replace it with an advantage-weighted α-divergence penalty while keeping ϵ as the trust-region radius (Section 3.1, Proposition 3.1).
2. We show that α sets the skewness of the penalty, i.e. which deviations from the target it resists most, and propose update rules that adapt α at every update from batch statistics, with a monotone improvement guarantee for any choice of α (Theorem 4.1).
3. On MiniGrid, MuJoCo with deep policies and Procgen, ALDIPPO keeps ratio deviations bounded as depth grows, is less sensitive to ϵ, and matches or improves on PPO and SPO.

2 Background

2.1 Reinforcement Learning

We consider a discounted Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \rho)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathbb{P}$ the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$ the reward function bounded by $R > 0$, $\gamma \in (0, 1)$ the discount factor, and ρ the initial state distribution (Sutton & Barto, 2018). A policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps states to distributions over actions. The goal of reinforcement learning is to find a policy that maximizes the expected cumulative discounted reward $J(\pi_\theta) = \mathbb{E}_{s_0 \sim \rho}[V^{\pi_\theta}(s_0)]$, where $V^\pi(s)$ is the value function under π.

Two useful related quantities are the *discounted state-occupancy measure* $d_\rho^{\pi_\theta}(s)$ and the *discounted state-action occupancy measure* $\nu_\rho^{\pi_\theta}(s, a)$, defined by equation 1:

$$d_\rho^{\pi_\theta}(s) = (1-\gamma)\,\mathbb{E}_{s_0 \sim \rho} \sum_{t \geq 0} \gamma^t \Pr(s_t = s \mid s_0, \pi_\theta), \qquad \nu_\rho^{\pi_\theta}(s, a) = d_\rho^{\pi_\theta}(s)\,\pi_\theta(a \mid s), \tag{1}$$

which capture the expected discounted visitation frequencies of states and state-action pairs under policy π_θ. Based on these quantities, Kakade & Langford (2002) derived the performance difference lemma, which lets us compute the change in expected return when switching from one policy to another. The performance difference may be expressed through equation 2:

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a) \sim \nu_\rho^{\pi'}}\big[A^\pi(s, a)\big], \tag{2}$$

where $A^\pi(s, a)$ is the advantage of action a in the state s under policy π. Replacing $\nu_\rho^{\pi'}$ with ν_ρ^{π} and reweighting actions by the importance ratio $w = \pi'/\pi$ gives the surrogate objective $L_\pi(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{(s,a) \sim \nu_\rho^{\pi}}[\,w\,A^\pi(s, a)]$. This surrogate is a first order approximation of equation 2, motivating the use of trust regions to avoid diverging estimates due to stale data (Schulman et al., 2015).

2.2 Proximal Policy Optimization

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is a policy gradient algorithm that uses a clipped surrogate objective to limit the policy updates, which may be defined as equation 3:

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_{(s,a) \sim \nu_\rho^{\pi_{\theta_{\mathrm{old}}}}}\Big[\min\big(w_\theta(s, a)\,\hat{A}(s, a), \mathrm{clip}(w_\theta(s, a), 1-\epsilon, 1+\epsilon)\,\hat{A}(s, a)\big)\Big], \tag{3}$$

where the advantage $\hat{A}(s,a)$ is an estimate of $A^{\pi_{\theta_{\text{old}}}}$ using Generalized Advantage Estimation (GAE) (Schulman et al., 2016). The clip parameterized by ϵ is motivated as a first-order alternative to the trust-region constraint in Trust Region Policy Optimization (TRPO) (Schulman et al., 2015): when $\hat{A} > 0$ the gradient is zero whenever $w_\theta(s,a) > 1+\epsilon$, and when $\hat{A} < 0$ the gradient is zero whenever $w_\theta(s,a) < 1-\epsilon$. The clipped objective is itself an alternative to a penalty-based update proposed in the same work (Schulman et al., 2017), which uses a KL divergence penalty:

$$L^{\text{KL}}(\theta) = \mathbb{E}_{(s,a)\sim\nu_\rho^{\pi_{\theta_{\text{old}}}}}\left[w_\theta \hat{A} - \beta\, D_{\text{KL}}[\pi_{\theta_{\text{old}}}(\cdot|s) \,\|\, \pi_\theta(\cdot|s)]\right], \tag{4}$$

with a heuristically adapted weight β between updates. In particular, Schulman et al. (2017) state that this heuristic β underperformed the clipped objective, motivating the use of the latter in practice.

2.3 Divergences

The KL divergence in equation 4 is the most popular choice of divergence, or measure of dissimilarity between distributions, in machine learning. However, it is one member of a family of divergences known as f-divergences (Csiszár, 1967), which more generally may be defined for π_θ, $\pi_{\theta_{\text{old}}}$ by:

$$D_f(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}})(s) = \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(\cdot|s)}\big[f(w_\theta(s,a))\big], \tag{5}$$

where f is a convex function with $f(1) = 0$ called the *generator* of D_f. The generator is defined up to a $c(w-1)$ term for any $c \in \mathbb{R}$, i.e. $D_f = D_{f+c(w-1)}$, and for KL $c = 1$. Due to its expectation form, it is also possible to estimate D_f from the batch collected from $\pi_{\theta_{\text{old}}}$. In this work, we focus on the α-divergence family (Amari, 2016), whose generator is:

$$f_\alpha(w) = \frac{w^\alpha - \alpha w + \alpha - 1}{\alpha(\alpha-1)}. \tag{6}$$

With slight abuse of notation, we denote the α-divergence between π_θ and $\pi_{\theta_{\text{old}}}$ by $D_\alpha(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}})$, as the rest of the paper will center on this particular family. This generator is normalized such that $f_\alpha(1) = f'_\alpha(1) = 0$ and $f''_\alpha(1) = 1$ for every α. Since equation 6 is undefined at $\alpha \in \{0,1\}$ the generator is defined by its limit, resulting in the KL divergence in equation 4 at $\alpha \to 0$ and its reverse at $\alpha \to 1$. The χ^2 divergence used by Xie et al. (2025) is recovered at $\alpha = 2$.

3 Revisiting PPO updates

3.1 PPO dynamics

The zero-gradient region of PPO's clipped objective is meant to limit policy updates to mitigate data staleness. However, this does not account for data batch interactions. Consider a simplified scenario where we update parameters θ with a batch that contains two samples x_1, x_2 with positive advantages, x_1 with $w > 1+\epsilon$, i.e. its gradient will be zero, and x_2 with $w < 1+\epsilon$. At a policy iteration step, the update for θ will have gradient contributions only from x_2, which will push the policy towards increasing w for x_2. Since both samples depend on the same θ, this update can then increase w for x_1 too, as conceptualized in the first panel of Figure 1.

In contrast, a penalty-based update has a non-zero gradient everywhere that points to a target ratio, resulting in gradient updates that don't drift arbitrarily far. SPO (Xie et al., 2025) weights the χ^2 penalty of each sample by its advantage, maximizing $\hat{A}w - \frac{|\hat{A}|}{\epsilon}\frac{(w-1)^2}{2}$ per sample. This sets a stationary point at $w^\star = 1 + \epsilon\,\operatorname{sgn}\hat{A}$, or in other words, at the clip boundary. The gradients resulting from this divergence then grow linearly with the distance from the target. We generalize this approach by considering the family of α-divergences, allowing to treat the steepness of the penalty as a design variable, as pictured in Figure 1. This motivates the generalized penalty objective:

$$\mathcal{L}_\pi(\theta) \;=\; -\,\mathbb{E}_{(s,a)\sim\nu_\rho^{\pi_{\theta_{\text{old}}}}}\left[\hat{A}(s,a)\, w_\theta(s,a) \;-\; \frac{|\hat{A}(s,a)|}{\epsilon}\, f_\alpha\big(w_\theta(s,a)\big)\right]. \tag{7}$$

Notably, due to the convexity of f_α, we may find a unique maximizer for each sample in equation 7 for a general α under mild assumptions, shown in Proposition 3.1:

Proposition 3.1 (Stationary ratio). *Let $\hat{A} \neq 0$ and $\epsilon|\alpha - 1| < 1$. The per-sample objective $\hat{A}w - \frac{|\hat{A}|}{\epsilon} f_\alpha(w)$ is strictly concave in w with unique maximizer $w^\star_\alpha$:*

$$w^\star_\alpha = \left(1 + (\alpha - 1)\,\epsilon \operatorname{sgn} \hat{A}\right)^{\frac{1}{\alpha - 1}}, \tag{8}$$

with $w^\star_2 = 1 \pm \epsilon$ and $w^\star_1 = e^{\pm\epsilon}$. (Proof in Appendix A.)

3.2 What does α do?

Since $f''_\alpha(1) = 1$ for every α, all α-divergences share similar curvature near $w = 1$. Each divergence differs from third-order terms in the Taylor expansion of f_α around $w = 1$:

$$f_\alpha(1 + \delta) = \frac{\delta^2}{2} + \sum_{k \geq 3} \frac{(\alpha - 2)(\alpha - 3) \cdots (\alpha - k + 1)}{k!} \delta^k = \frac{\delta^2}{2} + \frac{\alpha - 2}{6}\,\delta^3 + \mathcal{O}(\delta^4). \tag{9}$$

Just as the quadratic term describes local curvature, the cubic term defines the *skewness* of the penalty, or how curvature changes as we move away from the target (Amari, 2016), as visualized in Figure 1. At $\alpha = 2$, the cubic term is zero and thus the per-sample loss is symmetric around the target; with $\alpha > 2$ the term is positive and induces a steeper loss to the right of the target, while $\alpha < 2$ steepens the loss to its left. Intuitively, a higher α limits how far an update can raise the probability of actions whose advantage may be overestimated, while a smaller value prevents driving probabilities to zero.

Over training, we may prefer one skew of f_α over the other. Each policy update defines a new surrogate based on a collected batch, and prior work shows that the optimal hyperparameters can change as training progresses (Mohan et al., 2023). Further, optimization may degrade under this nonstationarity (Ellis et al., 2024). Additionally, advantage approximation is also subject to nonstationary targets (Igl et al., 2021), which may lead to poor estimates (Ilyas et al., 2020). We therefore choose α at every update from batch statistics, and propose update rules based on the following: how the batch is set to move the ratios, and how well its advantages may be trusted.

Batch direction Once a batch $\mathcal{B}$ is collected from $\pi_{\theta_{\text{old}}}$, each sample's estimated $\hat{A}$ defines if the ratios should go up ($\hat{A} > 0$) or down ($\hat{A} < 0$). We thus may choose α to minimize the penalty contribution in equation 7 for $\mathcal{B}$, which we define as $\Psi(\alpha) = \mathbb{E}_{\mathcal{B}}[|\hat{A}|\, f_\alpha(w)]$. However, this expectation has no closed form minimizer. To approximate it, we take the generator's Taylor expansion up to the fourth order and minimize Ψ over α, which yields the update rule $\alpha^{\mathcal{B}}$:

$$\alpha^{\mathcal{B}} = \frac{5}{2} - \frac{2M_3}{M_4}, \qquad M_k = \mathbb{E}_{\mathcal{B}}\left[|\hat{A}|\,(w - 1)^k\right], \tag{10}$$

where M_3 measures whether the batch mostly raises or lowers ratios and M_4 how far they spread. To preserve presentation clarity, we defer the theoretical derivation to Appendix B.

Since the ratios w are only known after the update, we derive three variants of $\alpha^{\mathcal{B}}$ using proxies for $w - 1$ built from the advantages. For instance, taking the targets of Proposition 3.1 to first order, $w - 1 \approx \epsilon \operatorname{sgn} \hat{A}$, yields $\alpha_t = \alpha_{\min} - 2\bar{A}/(\epsilon\,\overline{|A|})$. As the expansion only holds for small ratio changes, we use equation 10 as a direction rather than an exact optimum: the rules start from $\alpha_{\min}$ and raise α only when the batch is set to lower more ratios than it raises (Appendix B).

Advantage approximation quality We also consider an update rule based on the assumption that the advantage estimates may be poor or noisy, from the observation by Ilyas et al. (2020). When they are poor estimates, $\hat{A} = A + e$, the error e enters the surrogate as $\mathbb{E}[(w - 1)\, e]$, so its effect grows with how far the update moves the ratios. Following the intuition above, we then raise α to limit the influence of samples whose advantage may be overestimated. We estimate quality with two batch statistics, the SNR of the advantages and the explained variance of the critic:

$$q^{\text{SNR}} = \frac{\bar{A}^2}{\bar{A}^2 + \sigma_A^2}, \qquad q^{\text{EV}} = 1 - \frac{\operatorname{Var}(G - V)}{\operatorname{Var}(G)}, \tag{11}$$

where G are empirical returns and V the critic's values. The update rule α^{SNR} then combines smoothed versions of both into a score $q_t \in [0, 1]$ and sets $\alpha^{\text{SNR}} = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})(1 - q_t)$. Again, to preserve presentation clarity, we defer the derivation to Appendix B.2.

We summarize our update rules as follows:

Remark 1 — Update rules
Every update rule returns one clamped scalar per update from the default $\alpha_0 = \alpha_{\min}$, with EMA smoothing across updates where applicable. Here $\sigma(\cdot)$ is the logistic sigmoid.

$$\alpha^{\mathcal{B}1}: \quad \alpha_t = \alpha_{\min} - \frac{2\,\bar{A}}{\epsilon\,\overline{|A|}}, \tag{12a}$$

$$\alpha^{\mathcal{B}2}: \quad \alpha_t = \mathrm{EMA}_{0.9}\Big[\mathrm{clip}\Big(\alpha_{\min} - 2M_3^{\tanh}/M_4^{\tanh}\Big)\Big], \qquad \delta_i = \epsilon\tanh(\hat{A}_i/\sigma_A), \tag{12b}$$

$$\alpha^{\mathcal{B}3}: \quad \alpha_t = \mathrm{EMA}_{0.9}\Big[\mathrm{clip}\Big(\alpha_{\min} - 2g_1/(\epsilon\, g_2)\Big)\Big], \tag{12c}$$

$$\alpha^{\mathrm{SNR}}: \quad \alpha_t = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})(1 - q_t), \tag{12d}$$

$$q_t = \mathrm{EMA}_{0.9}\Big[\sqrt{\sigma\big(20(\tfrac{|\bar{A}|}{\sigma_A} - 0.1)\big)\; e^{-\mathrm{Var}(G-V)/\mathrm{Var}(G)}}\Big].$$

4 Theoretical guarantees

We give two guarantees for the objective equation 7. Section 4.1 shows that exact updates improve the return for any α, and Section 4.2 bounds the number of trajectories the sampled update needs.

4.1 Monotone improvement of the α-divergence regularized objective

Let $\bar{D}_\alpha = \mathbb{E}_{s\sim d_\rho^{\pi_{\theta_{\mathrm{old}}}}}[D_\alpha(\pi_\theta \,\|\, \pi_{\theta_{\mathrm{old}}})(s)]$ be the average α-divergence, with $D_\alpha(\pi_\theta \,\|\, \pi_{\theta_{\mathrm{old}}})(s)$ as in equation 5, and let $\xi = \max_s |\mathbb{E}_{a\sim\pi_\theta}[A^{\pi_{\theta_{\mathrm{old}}}}(s,a)]|$ (Achiam et al., 2017).

Theorem 4.1 (Monotone improvement for every α)**.** *Let* $m_\alpha = \min(w_{\min}^{\alpha-2},\, w_{\max}^{\alpha-2})$, $C_\alpha = \frac{2\gamma\,\xi}{(1-\gamma)^2\sqrt{2m_\alpha}}$ *and* $c_{\max} = \max(2-\alpha_{\min},\, \alpha_{\max}-2)$.

(a) If the ratios lie in $[w_{\min}, w_{\max}] \ni 1$ *(Assumption 1, Appendix C), then for all* θ *and* α,

$$J(\pi_\theta) \;\geq\; J(\pi_{\theta_{\mathrm{old}}}) \;+\; L_{\pi_{\theta_{\mathrm{old}}}}(\pi_\theta) \;-\; C_\alpha\sqrt{\bar{D}_\alpha}. \tag{13}$$

Consequently, if π_{t+1} *maximizes the right-hand side of equation 13 with* $\pi_{\theta_{\mathrm{old}}} = \pi_t$ *and any* $\alpha_t \in [\alpha_{\min}, \alpha_{\max}]$*, then* $J(\pi_{t+1}) \geq J(\pi_t)$ *for every t, and* $J(\pi_t)$ *converges.*

(b) If the ratios lie in the interval $W_{\alpha,\epsilon} = [w_-^\star, w_+^\star]$ *between the stationary ratios equation 8 (Assumption 2) and* $\epsilon\,|\alpha-1| \leq \eta_0 < 1$*, let* $r_\alpha = \max\big(w_+^\star - 1,\; 1 - w_-^\star\big)$ *and* $A_{\max} = \max_s \mathbb{E}_{a\sim\pi_{\theta_{\mathrm{old}}}}\big|A^{\pi_{\theta_{\mathrm{old}}}}(s,a)\big|$*. Then*

$$\big|J(\pi_\theta) - J(\pi_{\theta_{\mathrm{old}}}) - L_{\pi_{\theta_{\mathrm{old}}}}(\pi_\theta)\big| \;\leq\; \frac{\gamma A_{\max}}{(1-\gamma)^2}\, r_\alpha^2, \qquad r_\alpha = \epsilon + \frac{|\alpha-2|}{2}\,\epsilon^2 + \mathcal{O}(\epsilon^3), \tag{14}$$

with $r_2 = \epsilon$*. For* $\alpha \in [\alpha_{\min}, \alpha_{\max}]$*, this bound exceeds its value at* $\alpha = 2$ *by at most* $\frac{\gamma A_{\max}}{(1-\gamma)^2}\, c_{\max}\,\epsilon^3 + \mathcal{O}(\epsilon^4)$*. (Proof in Appendix C.)*

Part (a) gives a lower bound on the return for every α, so the minorization–maximization argument of TRPO (Schulman et al., 2015) applies to the whole family. By the AM–GM inequality, maximizing $L_{\pi_{\theta_{\mathrm{old}}}}(\pi_\theta) - \frac{\bar{a}}{2\epsilon}\bar{D}_\alpha$ with $\bar{a} = \mathbb{E}|\hat{A}|$, a penalized objective of the same form as equation 7, maximizes this lower bound up to an additive term of order ϵ (Lemma C.4). Appendix A.1 explains the per-sample weight $|\hat{A}|$ in equation 7.

For fixed $w_{\min}$ and $w_{\max}$, C_α is smallest at $\alpha = 2$, since $m_\alpha \leq m_2 = 1$. Part (b) shows that this comes from fixing the interval: on $W_{\alpha,\epsilon}$, the leading term $\gamma A_{\max}\epsilon^2/(1-\gamma)^2$ of equation 14 is the same for every α and matches the bound of TRPO (Schulman et al., 2015, Theorem 1) with total variation distance $\epsilon/2$. Part (a) holds for any sequence α_t, including a constant one, so it shows that changing α keeps the guarantee, but not that it improves the return, which we test in Section 5.

4.2 GLOBAL CONVERGENCE AND SAMPLE COMPLEXITY

We now analyze an algorithm that, like PPO, samples a new batch at every iteration and takes several gradient steps on it. As in Lemma C.4, we replace the weight $|\hat{A}|$ in equation 7 by a constant $a_0 > 0$.

Assumptions. We use the general assumptions of Appendix D.1: the score $\nabla_\theta \log \pi_\theta$ is bounded by G and B-Lipschitz (Assumption 3), the Fisher matrix satisfies $F_\rho(\theta) \succeq \mu_F I$, and the transferred compatible-function-approximation error is at most ϵ_{bias} (Assumption 4). In addition, we assume that the Hessian of the log-policy is B_2-Lipschitz (Assumption 5). This is reasonable in our setting, since in the overparameterized regime the Hessian of the log-policy is smooth for networks with smooth activations, such as the $\tanh$ networks we use on MiniGrid and MuJoCo.

Algorithm. At iteration k, we sample N pairs $(s_i, a_i) \sim \nu_\rho^{\pi_{\theta_k}}$ with unbiased advantage estimates $\hat{A}_i$, using the sampler of Agarwal et al. (2021, Algorithm 3), and define

$$\hat{\theta}_{k+1} \;=\; \arg \max_{\|\theta' - \theta_k\| \le c_r \eta} \; \frac{1}{N} \sum_{i=1}^{N} \Big[\hat{A}_i \, w_i(\theta') \;-\; \frac{a_0}{\epsilon} \, f_\alpha\big(w_i(\theta')\big) \Big], \qquad w_i(\theta') = \frac{\pi_{\theta'}(a_i|s_i)}{\pi_{\theta_k}(a_i|s_i)}. \tag{15}$$

We take T steps of projected gradient ascent on equation 15 from θ_k, with the step size h_{in} of Lemma D.4, and set θ_{k+1} to the last iterate, not the exact maximizer. Here $\eta = \epsilon(1-\gamma)/a_0$ is the step size of the natural policy gradient step that the update approximates, and $c_r = 4GR/((1-\gamma)^2 \mu_F)$ makes the ball contain the maximizer of the population objective (Lemma D.2). The projection keeps every ratio in $[1 - c_w\epsilon,\, 1 + c_w\epsilon]$ (Appendix D.1); PPO does not use it.

Analysis. Under Assumptions 3 and 4, Lemma D.1 bounds the optimality gap by the error of the update direction from the natural gradient $F_\rho(\theta_k)^{-1} \nabla J(\theta_k)$. Since $f''_\alpha(1) = 1$, the penalty has Hessian $(a_0/\epsilon) F_\rho(\theta_k)$ at θ_k for every α, so the population objective is strongly concave on the ball (Lemma D.2). Assumption 5 enters the bounds only through Lemma D.3: it bounds the change of the batch Hessian between the batch and population maximizers, so the batch maximizer has conditional bias $\mathcal{O}(1/N)$. Without it, our argument gives $\mathcal{O}(1/\sqrt{N})$ and a complexity of $\mathcal{O}(\bar{\varepsilon}^{-4})$.

In Theorem 4.2, we set $a_0 = \Theta(R/(1-\gamma))$, the scale of the advantages, and $\mathcal{O}(\cdot)$ hides polynomial factors in G, B, B_2 and $1/\mu_F$; the exact parameters are given in equation 87.

Theorem 4.2 (Global convergence and sample complexity)**.** *Suppose Assumptions 3–5 hold, let π^* be an optimal policy with $J^* = J(\pi^*)$ and $\mathrm{KL}_0 = \mathbb{E}_{s\sim d_\rho^{\pi^*}}[D_{\mathrm{KL}}(\pi^*(\cdot|s) \,\|\, \pi_{\theta_0}(\cdot|s))] < \infty$, and let $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. For $0 < \bar{\varepsilon} \le \bar{\varepsilon}_0$, with $\bar{\varepsilon}_0$ given in equation 88, choose η, N, K and T as in equation 87, so that $\epsilon = \mathcal{O}\big(\frac{(1-\gamma)^2\,\bar{\varepsilon}}{(1+|\alpha-2|)\,R}\big)$, $N = \mathcal{O}\big(\frac{(1+|\alpha-2|)\,R}{(1-\gamma)^2\,\bar{\varepsilon}}\big)$, $K = \mathcal{O}\big(\frac{\mathrm{KL}_0\,(1+|\alpha-2|)\,R^2}{(1-\gamma)^4\,\bar{\varepsilon}^2}\big)$ and $T = \mathcal{O}\big(\log \frac{R}{(1-\gamma)\,\bar{\varepsilon}}\big)$. Then*

$$J^* - \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\big[J(\theta_k)\big] \;\le\; \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \bar{\varepsilon},$$

and the algorithm samples $2NK = \mathcal{O}\big(\frac{\mathrm{KL}_0\,(1+|\alpha-2|)^2\,R^3}{(1-\gamma)^6\,\bar{\varepsilon}^3}\big)$ trajectories. (Proof in Appendix D.)

The curvature of the penalty gives the rate: Lemma D.1 charges the update direction at first order only through its conditional mean, so the $\mathcal{O}(1/N)$ bias of the batch maximizer, rather than its $\mathcal{O}(1/\sqrt{N})$ fluctuation, sets $N = \mathcal{O}(\bar{\varepsilon}^{-1})$. The argument fails for the clipped objective equation 3, which is constant past the clip boundary. For $T = 1$, the update is a stochastic policy gradient step, as $f'_\alpha(1) = 0$, with the $\mathcal{O}(\bar{\varepsilon}^{-4})$ bound of Liu et al. (2020, Theorem 4.6). The theorem also holds for $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ chosen from earlier batches, with $c_{\max}$ in place of $|\alpha - 2|$ (Appendix D).

5 EXPERIMENTS

5.1 MINIGRID ABLATIONS

To gauge the effectiveness of our update rules, we run two ablations on the MiniGrid benchmark (Chevalier-Boisvert et al., 2023) on the sparse-reward environment DoorKey-6x6. We compare

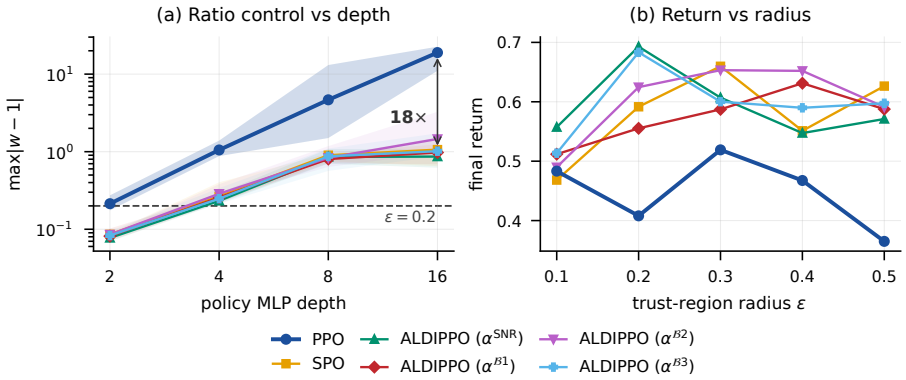

Figure 2: Ablation on MiniGrid DoorKey-6x6 over 10 seeds. (a) Ratio deviation $\max|w-1|$ against policy MLP depth at $\epsilon = 0.2$. Lines show medians and bands show 95% bootstrap confidence intervals, the dashed line marks ϵ. (b) Median final return against ϵ at depth 2.

clipped PPO and SPO with ALDIPPO under each $\alpha^{\mathcal{B}}$ and α^{SNR}, training with each method for 500k steps over 10 seeds, the results are visualized in Figure 2.

First, we vary the policy MLP depth from 2 to 16 hidden layers at $\epsilon = 0.2$ to measure how each algorithm scales with network depth. We report the peak ratio deviation $\max|w-1|$ over training in Figure 2(a). We can see that clipped PPO already reaches the clip boundary at depth 2, growing to a ratio deviation of 19 with the deepest networks. In contrast, ALDIPPO update rules stay bounded and plateau between 0.87 and 1.46, with SPO following a similar trend. In Figures 2(b) we sweep ϵ from 0.1 to 0.5 at depth 2. A larger ϵ allows for bigger updates, i.e. more data staleness, and makes clipping less reliable. Clip PPO is erratic, showing its sensitivity to ϵ, and reaches its lowest final return of 0.37 at $\epsilon = 0.5$, while ALDIPPO and SPO show more resilience to the variations of ϵ.

5.2 MuJoCo Gym

We evaluate the performance of ALDIPPO with deep policies on eight representative MuJoCo Gym environments. We train every method for ten million environment steps over six seeds. To compare our results to SPO (Xie et al., 2025), we use MLPs with 256 hidden dimensions and 7 layers. The details of training hyperparameters are in Appendix F and results with 2-layer policies are in Appendix F.1. Figure 3 plots episodic returns over environment steps and Table 1 shows final returns, where we note $\alpha^{\mathcal{B}3}$ ALDIPPO has the best aggregate performance. Particularly at 7 layers, where Xie et al. (2025) report PPO collapsing, PPO is the worst method on most environments.

Table 1: MuJoCo return over the last 10% of training, reported as mean with standard deviation in parentheses over six seeds at 10M steps. Bold marks the best method in each environment. The final row scales each environment so the worst method is zero and the best is one and reports the median over the eight environments. †HumanoidStandup is reported in thousands.

Environment	PPO	SPO	α^{SNR}	$\alpha^{\mathcal{B}1}$	$\alpha^{\mathcal{B}2}$	$\alpha^{\mathcal{B}3}$
HumanoidStandup†	57.8 (10.3)	**171 (22)**	156 (3)	154 (3)	158 (3)	159 (3)
Walker2d	1105 (500)	3350 (248)	3246 (554)	3942 (960)	3612 (476)	**3967 (347)**
Humanoid	283 (57)	2083 (561)	2011 (261)	2049 (620)	1747 (431)	**2302 (442)**
Ant	−58.7 (23.4)	785 (508)	982 (605)	365 (488)	737 (435)	**1021 (325)**
Pusher	−334 (6)	−114 (47)	**−70.4 (6.3)**	−205 (65)	−210 (73)	−230 (54)
Hopper	937 (186)	1522 (260)	1466 (187)	**1685 (269)**	1673 (213)	1644 (176)
HalfCheetah	1439 (812)	4547 (1368)	4619 (1332)	**4623 (1502)**	4422 (1576)	4382 (1908)
Swimmer	90.3 (23.6)	**103 (18)**	69.7 (28.6)	92.2 (25.3)	89.2 (26.4)	97.9 (24.1)
Median normalised	0.000	0.864	0.862	0.863	0.807	**0.934**

5.3 Procgen

Finally, we evaluate the methods with $\epsilon = 0.5$ on the Procgen benchmark by Cobbe et al. (2020) in Figure 3 and Table 2; experimental results with $\epsilon = 0.2$ on all sixteen games and details on the

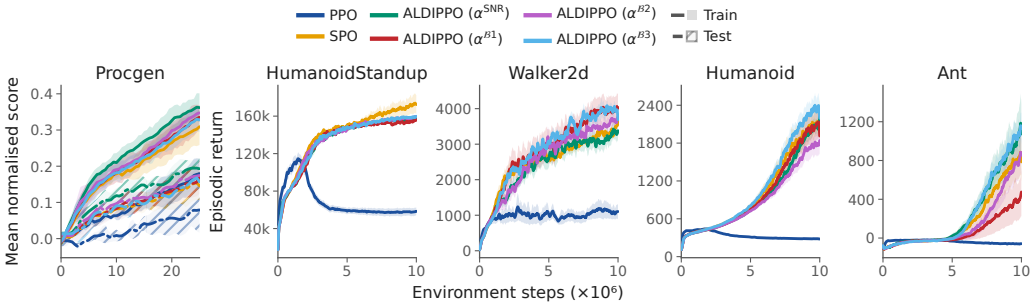

Figure 3: Left: Procgen mean normalised score at $\epsilon = 0.5$ over eight games, reported as the mean over three seeds with one standard deviation. Remaining panels: MuJoCo Gym episodic return on four representative environments, reported as the mean over six seeds with one standard error bands. All eight MuJoCo environments are in Figure 8 and the per-game Procgen curves are in Figure 7.

chosen training hyperparameters are in Appendix E, where we can see that all ALDIPPO update rules outperform clip PPO in aggregate.

For the high ϵ tests, we train on eight games in easy mode and test on the full level distribution. Each run uses 200 training levels and 25M environment steps. We report three seeds per algorithm, normalizing returns with the constants from Cobbe et al. (2020). We note that α^{SNR} requires a lower ceiling $\alpha_{\max} = 2$ and retuned gate constants at this ϵ (Appendix E); the remaining update rules use the same hyperparameters as at $\epsilon = 0.2$. We note that ALDIPPO degrades more softly between $\epsilon = 0.5$ and $\epsilon = 0.2$ compared to PPO, which takes a sharp performance decline. We also note that SPO has the lowest aggregate performance at both $\epsilon = 0.2$ and $\epsilon = 0.5$ among penalty methods. We hypothesize that by being able to adapt the penalty to the observed batch, ALDIPPO is able to generalize across levels better than SPO.

Table 2: Procgen held-out test return at $\epsilon = 0.5$, reported as mean with standard deviation in parentheses over three seeds and over 25M steps on the full level distribution. Bold marks the best method in each game. The final row is the mean normalised score over the eight games, using the easy-mode constants from Cobbe et al. (2020). The α^{SNR} column uses the update rule ceiling $\alpha_{\max} = 2$ discussed in Appendix E; the remaining columns use the same hyperparameters as the $\epsilon = 0.2$ runs.

Environment	PPO	SPO	α^{SNR}	$\alpha^{\mathcal{B}1}$	$\alpha^{\mathcal{B}2}$	$\alpha^{\mathcal{B}3}$
BigFish	1.0 (.0)	2.4 (1.2)	2.6 (1.3)	**2.9 (1.8)**	2.2 (1.4)	2.4 (1.1)
BossFight	6.5 (.9)	**8.1 (.4)**	7.9 (.3)	7.8 (.4)	7.7 (.6)	7.2 (.3)
CaveFlyer	2.9 (1.2)	4.0 (1.2)	**4.8 (.7)**	4.5 (1.1)	4.5 (.5)	4.3 (.6)
Chaser	3.4 (1.8)	4.9 (2.2)	5.8 (.7)	5.1 (2.2)	**6.7 (.5)**	5.9 (1.0)
Climber	2.4 (.4)	3.6 (.3)	3.6 (.2)	3.4 (.3)	3.6 (.4)	**3.7 (.0)**
Dodgeball	.6 (.2)	1.4 (.1)	1.3 (.2)	1.3 (.2)	**1.4 (.1)**	1.3 (.1)
Leaper	3.3 (1.6)	3.0 (1.0)	**3.8 (1.4)**	3.0 (1.1)	2.9 (1.3)	3.2 (1.3)
Plunder	3.6 (.5)	4.8 (1.1)	**6.7 (.6)**	6.3 (.7)	5.5 (1.2)	5.4 (1.4)
Mean normalised	.078 (.061)	.150 (.043)	**.194 (.028)**	.163 (.042)	.174 (.020)	.163 (.025)

6 Related work

Trust regions and clipping. Bounding policy updates to mitigate data staleness goes back to conservative policy iteration (Kakade & Langford, 2002) and TRPO (Schulman et al., 2015). PPO (Schulman et al., 2017) replaced this constraint with a clipped surrogate objective. Wang et al. (2020) observe that the clip does not bound the ratio and add a rollback term past the clip boundary. SPO (Xie et al., 2025) replaces the clip with an advantage-weighted χ^2 penalty, which ALDIPPO recovers at $\alpha = 2$. Mirror descent (Tomar et al., 2022) and MPO (Abdolmaleki et al., 2018) also regularize each update with a divergence, but use the KL divergence.

f- and α-divergences in reinforcement learning. Belousov & Peters (2017; 2019) derive policy updates under general f-divergence constraints. Tsallis entropy, which is closely related to α-divergences, has been used as a policy regularizer rather than as a trust region (Lee et al., 2018; 2019). Brekelmans et al. (2022) analyze α-divergence regularizers and their connection to Tsallis and Rényi entropies in finite tabular MDPs. The closest work to ours is Xu et al. (2023), who add an α-divergence penalty with a fixed α to PPO; their analysis is limited to diagonal Gaussian policies. ALDIPPO penalizes the importance ratio of each sample, so it applies to any policy class, with discrete or continuous actions, and it adapts α at every update.

Divergences in preference optimization. Language-model alignment has recently revisited alternative divergences. Go et al. (2023) fine-tune towards a target distribution by minimizing f-divergences. Wang et al. (2024) generalize DPO to f-divergence constraints. Huang et al. (2024) show that a χ^2 penalty makes direct alignment pessimistic and mitigates overoptimization. These methods regularize towards a fixed reference model, while ALDIPPO regularizes each update towards the policy that collected the batch.

Adapting hyperparameters during training. The best hyperparameters change as training progresses (Mohan et al., 2023), partly because both the data and the value targets are nonstationary (Igl et al., 2021; Ellis et al., 2024). The KL-penalty variant of PPO adapts the penalty weight β between updates from the measured KL divergence (Schulman et al., 2017). Our update rules keep the penalty scale fixed through ϵ and adapt the shape of the penalty from batch statistics instead.

Sample complexity. Agarwal et al. (2021) prove global convergence of policy gradient methods, including natural policy gradient with function approximation, and our analysis follows the setting of Liu et al. (2020). Stochastic policy gradient needs $\mathcal{O}(\bar{\varepsilon}^{-4})$ trajectories, and variance-reduced policy gradient needs $\mathcal{O}(\bar{\varepsilon}^{-3})$ under an assumption on the variance of the importance weights (Liu et al., 2020). Accelerated natural policy gradient needs $\mathcal{O}(\bar{\varepsilon}^{-2})$, using an inner loop that solves a regression problem for the natural gradient (Mondal & Aggarwal, 2024). Theorem 4.2 gives $\mathcal{O}(\bar{\varepsilon}^{-3})$ for a multi-epoch update of the kind PPO uses, without an assumption on the importance weights.

7 Conclusion

We revisited PPO's clipped surrogate and showed that, because samples share parameters, the clip objective need not keep the importance ratios near the trust region, resulting in PPO's sensitivity to the clip hyperparameter and network depth. ALDIPPO replaces the clip with an advantage-weighted α-divergence penalty with a gradient that points toward a target ratio set by ϵ, where α defines how each sample influences the update. We proposed update rules that choose α at every update based on how the batch is set to move the importance ratios and from how well calibrated its advantage estimates are. The monotone improvement and sample complexity guarantees hold for any α within the bounds of these update ruls. Empirically, ALDIPPO keeps ratio deviations bounded as depth grows on MiniGrid, outperforms PPO with deep policies on MuJoCo, and outperforms PPO in Procgen in regular ϵ settings, while showing lower degradation when ϵ increases.

Limitations and future work. The update rules estimate the ratios before the update from advantage-based proxies, and use the Taylor approximation of Section 3.1 as an update direction. α^{SNR} also required retuning at $\epsilon = 0.5$, indicating that other update rules may better accommodate for optimization in RL. Natural extensions are choosing α per state or per sample, and applying the penalty to the RL fine-tuning of language models, where PPO-style clipping is standard.

AI USE STATEMENT

In this work, we used generative AI tools to give feedback on and polish text written by the authors, search for related work, check and typeset derivations in LaTeX, and generate code for visualization. The main training code was developed by the authors, from which experiments were tested with LLMs. The authors manually verified that the generated code correctly implements the proposed experiments. The authors manually checked the references given by the generative model to verify the publication information is correct. All derivations, experiments and claims were verified by the authors, who take full responsibility for the content of the paper. No experimental results were generated by a language model.

REPRODUCIBILITY STATEMENT

All theoretical claims are proven in Appendices A–D with associated assumptions. All experimental details are specified in Section 5 and Appendices F–E. To this submission, we attach an anonymized zip file with the necessary code to replicate our results. We commit to releasing an open-source code repository at publication acceptance.

ETHICS STATEMENT

This work proposes a general-purpose policy optimization method evaluated on simulated benchmarks, and involves no human subjects or personal data. Improvements to policy optimization may be applied to downstream systems such as the fine-tuning of language models, but we are not aware of risks specific to our method beyond those of RL in general.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations (ICLR)*, 2018.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22:1–76, 2021.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, 2016. ISBN 978-4-431-56743-1. doi: 10.1007/978-4-431-55978-8.

Boris Belousov and Jan Peters. f-divergence constrained policy improvement. *arXiv preprint arXiv:1801.00056*, 2017.

Boris Belousov and Jan Peters. Entropic regularization of markov decision processes. *Entropy*, 21(7), 2019. ISSN 1099-4300. doi: 10.3390/e21070674. URL https://www.mdpi.com/1099-4300/21/7/674.

Rob Brekelmans, Tim Genewein, Jordi Grau-Moya, Grégoire Delétang, Markus Kunesch, Shane Legg, and Pedro A Ortega. Your policy regularizer is secretly an adversary. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=berNQMTYWZ.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & Miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, pp. 2048–2056, 2020.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Benjamin Ellis, Matthew Thomas Jackson, Andrei Lupu, Alexander David Goldie, Mattie Fellows, Shimon Whiteson, and Jakob N. Foerster. Adam on local time: Addressing nonstationarity in RL with relative Adam timesteps. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems*, 2024.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations (ICLR)*, 2020.

Gustavo L. Gilardoni. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Transactions on Information Theory*, 56(11):5377–5386, 2010.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11546–11583. PMLR, 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *CoRR*, abs/2407.13399, 2024. doi: 10.48550/ARXIV.2407.13399. URL https://doi.org/10.48550/arXiv.2407.13399.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Qun8fv4qSby.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations (ICLR)*, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, 2002.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics Autom. Lett.*, 3(2): 1466–1473, 2018. doi: 10.1109/LRA.2018.2800085. URL https://doi.org/10.1109/LRA.2018.2800085.

Kyungjae Lee, Sungyub Kim, Sungbin Lim, Sungjoon Choi, and Songhwai Oh. Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning. *arXiv preprint arXiv:1902.00137*, 2019.

Yanli Liu, Kaiqing Zhang, Tamer Başar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Aditya Mohan, Carolin Benjamins, Konrad Wienecke, Alexander Dockhorn, and Marius Lindauer. AutoRL hyperparameter landscapes. In Aleksandra Faust, Roman Garnett, Colin White, Frank Hutter, and Jacob R. Gardner (eds.), *International Conference on Automated Machine Learning, 12-15 November 2023, Hasso Plattner Institute, Potsdam, Germany*, volume 224 of *Proceedings of Machine Learning Research*, pp. 13/1–27. PMLR, 2023.

Washim Uddin Mondal and Vaneet Aggarwal. Improved sample complexity analysis of natural policy gradient algorithm with general parameterization for infinite horizon discounted reward Markov decision processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024. arXiv:2310.11677.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91577-7. doi: 10.1007/978-3-319-91578-4.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03924-6.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pp. 1057–1063, 2000.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2cRzmWXK9N.

Yuhui Wang, Hao He, and Xiaoyang Tan. Truly proximal policy optimization. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 113–122. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/wang20b.html.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Haotian Xu, Zheng Yan, Junyu Xuan, Guangquan Zhang, and Jie Lu. Improving proximal policy optimization with alpha divergence. *Neurocomputing*, 534(C):94–105, 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.02.008. URL https://doi.org/10.1016/j.neucom.2023.02.008.

APPENDIX CONTENTS

A Stationary ratios **13**

A.1 The per-sample weight 14

B Choosing α at each update **14**

B.1 Batch direction 14

B.2 Advantage approximation quality 15

C Improvement guarantees: proofs **16**

C.1 Proof of Theorem 4.1(b) 17

D Proofs of the sample-complexity results of Section 4.2 **18**

D.1 Setting and standing assumptions 18

D.2 Outer-loop analysis 18

D.3 The proximal step is an inexact natural-gradient step 19

D.4 Conditional bias of the batch proximal point 22

D.5 Inner loop and assembly 30

E Procgen experiment details **34**

F MuJoCo experiment details **34**

F.1 Two-layer policies 34

A STATIONARY RATIOS

We restate Proposition 3.1 with the properties that Appendix B and Section 4 use.

Proposition A.1 (Stationary ratio)**.** *Let $\hat{A} \neq 0$ and $\epsilon|\alpha - 1| < 1$. The per-sample objective $\ell(w) = \hat{A}w - \frac{|\hat{A}|}{\epsilon} f_\alpha(w)$ is strictly concave on $w > 0$, and its unique maximizer $w^\star_\alpha$ in equation 8 depends on $\hat{A}$ only through $\operatorname{sgn} \hat{A}$. For both signs of $\hat{A}$, $w^\star_\alpha$ strictly decreases as α increases, and*

$$w^\star_\alpha = 1 + \epsilon \operatorname{sgn} \hat{A} + \frac{2-\alpha}{2}\,\epsilon^2 + \mathcal{O}(\epsilon^3). \tag{16}$$

For $\hat{A} > 0$, the gain $\ell(w^\star_\alpha) - \ell(1)$ also strictly decreases as α increases.

Proof. Write $s = \operatorname{sgn} \hat{A}$. Since $\ell''(w) = -\frac{|\hat{A}|}{\epsilon} w^{\alpha-2} < 0$, ℓ is strictly concave, and $\ell'(w) = 0$ reads

$$\hat{A} = \frac{|\hat{A}|}{\epsilon} \cdot \frac{w^{\alpha-1} - 1}{\alpha - 1} \quad \Longleftrightarrow \quad w^{\alpha-1} = 1 + (\alpha - 1)\,\epsilon\, s. \tag{17}$$

This equation has the unique positive solution equation 8 when $1+(\alpha-1)\epsilon s > 0$, which $\epsilon|\alpha-1| < 1$ ensures for both signs, and the solution depends on $\hat{A}$ only through s. The cases $\alpha = 1$ and $\alpha = 0$ follow from $(1 + \beta\epsilon s)^{1/\beta} \to e^{\epsilon s}$ as $\beta = \alpha - 1 \to 0$ and from direct substitution.

Monotonicity. For $s = +1$ and $\beta = \alpha - 1$, $\log w^\star_\alpha = \log(1 + \beta\epsilon)/\beta$ is the slope of the chord from 0 to β of the concave map $x \mapsto \log(1 + x\epsilon)$, so it strictly decreases in β. For $s = -1$, the identity $w^\star_-(\alpha) = 1/w^\star_+(2 - \alpha)$ gives the same conclusion.

First-order form. Since $\log w_\alpha^\star = \log\big(1+(\alpha-1)\epsilon s\big)/(\alpha-1) = \epsilon s - (\alpha-1)\epsilon^2/2 + \mathcal{O}(\epsilon^3)$, exponentiating gives equation 16.

Gain. At the stationary point, $w^\alpha = w\, w^{\alpha-1} = w\big(1+(\alpha-1)\epsilon s\big)$. Substituting this into f_α and into $\ell(w^\star)-\ell(1) = \hat{A}(w^\star-1) - \frac{|\hat{A}|}{\epsilon} f_\alpha(w^\star)$ gives

$$f_\alpha(w^\star) = \frac{w^\star(\epsilon s - 1)+1}{\alpha}, \qquad \ell(w^\star)-\ell(1) = |\hat{A}|\,\Phi_\alpha^{(s)}(\epsilon), \tag{18}$$

$$\Phi_\alpha^{(s)}(\epsilon) = \frac{\big(1+s(\alpha-1)\epsilon\big)^{\alpha/(\alpha-1)} - 1 - s\,\alpha\epsilon}{\alpha\,\epsilon} = \frac{\epsilon}{2} - s\,\frac{(\alpha-2)\,\epsilon^2}{6} + \mathcal{O}(\epsilon^3). \tag{19}$$

At $\alpha = 2$, $\Phi_2^{(\pm)}(\epsilon) = \epsilon/2$, the value for SPO. For $s=+1$, $\Phi_\alpha^{(+)}(\epsilon) = \max_{w>0}\,[\,w - 1 - f_\alpha(w)/\epsilon\,]$ with maximizer $w_\alpha^\star > 1$, and $\partial_\alpha f_\alpha(w) = (\log w)^3\, T_1(w;\alpha) > 0$ for $w > 1$ by equation 22. Hence, for $\alpha < \alpha'$, $\Phi_{\alpha'}^{(+)}(\epsilon) = w_{\alpha'}^\star - 1 - f_{\alpha'}(w_{\alpha'}^\star)/\epsilon < w_{\alpha'}^\star - 1 - f_\alpha(w_{\alpha'}^\star)/\epsilon \le \Phi_\alpha^{(+)}(\epsilon)$.

Gradient. The gradient of the per-sample loss with respect to θ is $-\big[\hat{A} - \frac{|\hat{A}|}{\epsilon} f'_\alpha(w)\big]\nabla_\theta w$. Since f'_α is strictly increasing, the bracket vanishes only at $w = w_\alpha^\star$, and its magnitude strictly increases with $|w - w_\alpha^\star|$ on each side of $w_\alpha^\star$; for $\alpha > 1$ it grows like $w^{\alpha-1}$ as $w \to \infty$. The clipped objective equation 3, in contrast, has zero derivative for every w past the clip boundary (Xie et al., 2025, Eq. 8). □

A.1 The per-sample weight

The objective equation 7 weights the penalty of each sample by $|\hat{A}|/\epsilon$, while Lemma C.4 and Theorem 4.2 use a constant weight. We use two properties of the weight $|\hat{A}|/\epsilon$. First, equation 7 is homogeneous of degree one in the advantages, so rescaling all $\hat{A}$ by $c > 0$ changes neither its maximizer nor its stationary ratios. Second, by equation 17 the stationary ratio depends on $\hat{A}$ only through its sign, so all samples with $\hat{A} > 0$ share the target $w_+^\star$ and all samples with $\hat{A} < 0$ share the target $w_-^\star$. Assumption 2 and the proxy of $\alpha^{\mathcal{B}1}$ in Appendix B.1 rely on this. With a weight $h(|\hat{A}|)/\epsilon$, the first-order condition becomes $f'_\alpha(w^\star) = \epsilon \operatorname{sgn}\hat{A}\,|\hat{A}|/h(|\hat{A}|)$, so the target depends on $|\hat{A}|$ unless h is linear.

B Choosing α at each update

In this section, we give the theoretical motivation for the update rules of Remark 1 and derive them. The value of α can be chosen either from the direction in which the batch moves the ratios (Appendix B.1) or from the reliability of its advantages (Appendix B.2). All four rules of Remark 1 take values in $[\alpha_{\min}, \alpha_{\max}]$, so Theorem 4.1 holds for each of them.

B.1 Batch direction

Once $\mathcal{B}$ is collected, the sign of each $\hat{A}$ decides whether the update raises or lowers the ratio of that sample. We choose the α whose penalty is smallest for these changes, that is, the minimizer of the penalty of the batch in equation 7, $\Psi(\alpha) = \mathbb{E}_{\mathcal{B}}[\,|\hat{A}|\,f_\alpha(w)\,]$. Let $M_k = \mathbb{E}_{\mathcal{B}}[\,|\hat{A}|\,(w-1)^k\,]$ and $\delta_{\max} = \max_{\mathcal{B}}|w-1|$.

Theorem B.1 (Optimal geometry and its Newton step)**.** *Let $\mathcal{B}$ be a batch with ratios $w > 0$ that contains a sample with $\hat{A} \neq 0$ and $w \neq 1$.*

- *(a) Ψ is strictly convex in α, so it has a unique minimizer on $[\alpha_{\min}, \alpha_{\max}]$. If $\mathcal{B}$ contains samples with $\hat{A} \neq 0$ on both sides of $w = 1$, Ψ also has a unique minimizer on $\mathbb{R}$; otherwise Ψ is monotone in α.*
- *(b) There is a constant $c > 0$, depending only on $[\alpha_{\min}, \alpha_{\max}]$, such that if $\delta_{\max} \le c$, a Newton step on Ψ from any $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$ gives*

$$\hat{\alpha} = \frac{5}{2} - \frac{2M_3}{M_4} + \mathcal{O}\Big(\delta_{\max}\Big(1 + \frac{|M_3|}{M_4}\Big)\Big). \tag{20}$$

Proof. (a) The second divided difference of $f(x) = w^x$ at the nodes $0, 1, \alpha$ is $f_\alpha(w)$, and $f''(x) = w^x(\log w)^2$, so the Hermite–Genocchi formula $f[x_0, x_1, x_2] = \int_0^1 \int_0^{1-s} f''\big(x_0 + s(x_1 - x_0) + t(x_2 - x_0)\big)\, dt\, ds$ gives

$$f_\alpha(w) = (\log w)^2 \int_0^1 \int_0^{1-s} w^{\,s+t\alpha}\, dt\, ds. \tag{21}$$

Differentiating under the integral sign gives, for $n \geq 0$,

$$\partial_\alpha^n f_\alpha(w) = (\log w)^{n+2}\, T_n(w;\alpha), \qquad T_n(w;\alpha) = \int_0^1 \int_0^{1-s} t^n\, w^{\,s+t\alpha}\, dt\, ds \;>\; 0. \tag{22}$$

For $n = 2$ the right-hand side is positive for every $w \neq 1$, so $\Psi'' > 0$. For $n = 1$, $\partial_\alpha f_\alpha(w)$ has the sign of $\log w$: samples with $w > 1$ favour a smaller α and samples with $w < 1$ favour a larger one, so the minimizer depends on the whole batch. As $\alpha \to \infty$, $T_1(w;\alpha) \to \infty$ for $w > 1$ and $T_1(w;\alpha) \to 0$ for $w < 1$, and the reverse holds as $\alpha \to -\infty$. Hence Ψ' changes sign on $\mathbb{R}$ if and only if $\mathcal{B}$ contains samples with $\hat{A} \neq 0$ on both sides of $w = 1$.

(b) The Newton step from α_0 is $\alpha_0 - \Psi'(\alpha_0)/\Psi''(\alpha_0)$. Taking the expectation of equation 9 gives

$$\Psi(\alpha) = \frac{M_2}{2} + \frac{\alpha - 2}{6}\, M_3 + \frac{(\alpha - 2)(\alpha - 3)}{24}\, M_4 + \rho(\alpha), \tag{23}$$

where ρ collects the terms with M_k, $k \geq 5$. The coefficients of equation 9 and their first two derivatives in α grow at most polynomially in k on $[\alpha_{\min}, \alpha_{\max}]$, and $|M_k| \leq \delta_{\max}^{k-4} M_4$, so $|\rho'(\alpha)|, |\rho''(\alpha)| \leq C\delta_{\max} M_4$ for $\delta_{\max} \leq \frac{1}{2}$. Without ρ, $\Psi'(\alpha) = M_3/6 + (2\alpha - 5)M_4/24$ is affine and $\Psi'' = M_4/12$, so the Newton step from any α_0 gives the minimizer $\frac{5}{2} - 2M_3/M_4$ of the quadratic model. Let a and $b = M_4/12$ be the values of $\Psi'(\alpha_0)$ and $\Psi''(\alpha_0)$ without ρ. The Newton step differs from $\frac{5}{2} - 2M_3/M_4$ by $\big(a\rho''(\alpha_0) - b\rho'(\alpha_0)\big)/\big(b\,(b + \rho''(\alpha_0))\big)$. Since $|a|/b \leq 2|M_3|/M_4 + |\alpha_0 - \frac{5}{2}|$, this difference is of the order in equation 20 once $C\delta_{\max} \leq \frac{1}{24}$. □

From the Newton step to the update rules. The rules keep only the term $-2M_3/M_4$ of equation 20 and start from $\alpha_{\min}$ instead of $\frac{5}{2}$, so a batch with $M_3 = 0$ keeps $\alpha = \alpha_{\min}$, and α rises above $\alpha_{\min}$ only when $M_3 < 0$, that is, when the batch lowers more ratios than it raises. All rules are clipped to $[\alpha_{\min}, \alpha_{\max}]$.

Proxies for the ratios. The ratios are known only after the update, so each rule replaces $\delta_i = w_i - 1$ by a proxy computed from the advantages. All three proxies are invariant to the scale of the advantages.

- $\alpha^{\mathcal{B}1}$ uses the first-order target $\delta_i = \epsilon \operatorname{sgn} \hat{A}_i$ of equation 16, which is exact at $\alpha = 2$ and the same for all samples of the same sign (Appendix A.1). Then $M_3 = \epsilon^3\, \mathbb{E}_{\mathcal{B}}[\hat{A}]$ and $M_4 = \epsilon^4\, \mathbb{E}_{\mathcal{B}}|\hat{A}|$, which gives equation 12a. Normalizing the advantages per batch sets $\mathbb{E}_{\mathcal{B}}[\hat{A}] = 0$ and hence $\alpha = \alpha_{\min}$, so this rule uses unnormalized advantages.
- $\alpha^{\mathcal{B}2}$ uses $\delta_i = \epsilon \tanh(\hat{A}_i/\sigma_A)$. Like the targets, it satisfies $|\delta_i| \leq \epsilon$, but it also depends on the size of $\hat{A}_i$, so it stays informative after normalization. The batch moments give equation 12b, and an EMA across updates ($\beta = 0.9$) reduces the noise of M_3.
- $\alpha^{\mathcal{B}3}$ uses the linear proxy $\delta_i = \epsilon(\hat{A}_i - \bar{A})/\sigma_A$ with unweighted moments, so that $2M_3/M_4 = 2g_1/(\epsilon g_2)$, where g_1 and g_2 are the skewness and kurtosis of the advantages, which gives equation 12c. We drop the weight $|\hat{A}|$ because, for centred advantages, the weighted ratio is $\frac{\sigma_A}{\epsilon}\mathbb{E}[\operatorname{sgn}(\hat{A})|\hat{A}|^4]/\mathbb{E}[|\hat{A}|^5]$, a ratio of fourth- and fifth-order moments that is noisy at RL batch sizes. This rule also uses the EMA.

B.2 Advantage approximation quality

The rules of Appendix B.1 treat the advantages as exact. To model the estimation error, let $\hat{A} = A + e$, where e has zero mean given (s, a), and let $\sigma_e^2 = \mathbb{E}_{\mathcal{B}}[e^2]$. Then

$$\mathbb{E}_{\mathcal{B}}[w\hat{A}] = \mathbb{E}_{\mathcal{B}}[wA] + \mathbb{E}_{\mathcal{B}}[(w-1)\,e] + \mathbb{E}_{\mathcal{B}}[e], \qquad \big|\mathbb{E}_{\mathcal{B}}[(w-1)\,e]\big| \leq \sigma_e \left(\mathbb{E}_{\mathcal{B}}[(w-1)^2]\right)^{1/2}, \tag{24}$$

where the last term does not depend on θ and the bound is the Cauchy–Schwarz inequality. The update maximizes an objective built on $\mathbb{E}_{\mathcal{B}}[w\hat{A}]$, the return improves with $\mathbb{E}_{\mathcal{B}}[wA]$, and their difference grows with how far the ratios move. For samples with $\hat{A} > 0$, whose probability the update raises, a larger α reduces this movement but also the gain of the update (Proposition A.1). The next proposition shows how the best α on these samples depends on the error.

Proposition B.2. *Suppose that $\mathcal{B}$ contains samples with $\hat{A} > 0$ and that each of them moves to its target $w^\star_\alpha$. Let $m(\alpha) = \big(\mathbb{E}_{\mathcal{B}}[(w^\star_\alpha - 1)^2\,\mathbb{I}\{\hat{A} > 0\}]\big)^{1/2}$ and $G(\alpha) = \mathbb{E}_{\mathcal{B}}[\hat{A}\,\mathbb{I}\{\hat{A} > 0\}]\,\Phi^{(+)}_\alpha(\epsilon)$ be the movement and the gain on these samples. For $\sigma \geq 0$, let $\alpha^\star(\sigma)$ be a maximizer of $G(\alpha) - \sigma\, m(\alpha)$ over $[\alpha_{\min}, \alpha_{\max}]$, the gain minus the bound equation 24 on the error term with $\sigma_e = \sigma$. Then $\sigma < \sigma'$ implies $\alpha^\star(\sigma) \leq \alpha^\star(\sigma')$.*

Proof. G and m are continuous, so the maximizers exist. Let $\alpha = \alpha^\star(\sigma)$ and $\alpha' = \alpha^\star(\sigma')$. Adding $G(\alpha) - \sigma m(\alpha) \geq G(\alpha') - \sigma m(\alpha')$ and $G(\alpha') - \sigma' m(\alpha') \geq G(\alpha) - \sigma' m(\alpha)$ gives $(\sigma' - \sigma)\big(m(\alpha) - m(\alpha')\big) \geq 0$, so $m(\alpha) \geq m(\alpha')$. Since the target $w^\star_\alpha > 1$ strictly decreases in α (Proposition A.1), m is strictly decreasing, and hence $\alpha \leq \alpha'$. □

Since G also strictly decreases in α (Proposition A.1), $\alpha^\star(0) = \alpha_{\min}$, and $\alpha^\star(\sigma) = \alpha_{\max}$ for large σ: a smaller α gains more, and a larger α is less exposed to the error. On samples with $\hat{A} < 0$, a larger α increases the movement, which is the other side of the trade-off discussed after equation 9.

Measuring the error. The rule α^{SNR} in equation 12d follows Proposition B.2: it sets $\alpha = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})(1 - q_t)$, which increases as the estimated reliability $q_t \in [0, 1]$ decreases. In equation 11, $1 - q^{\mathrm{EV}}$ is the fraction of the variance of the returns that the critic, the baseline of $\hat{A}$, does not explain, and $1 - q^{\mathrm{SNR}}$ measures how much the spread of the advantages dominates their mean. In q_t, the SNR term enters through a sigmoid of $|\bar{A}|/\sigma_A$ centred at 0.1, and the EV term as $e^{-\operatorname{Var}(G-V)/\operatorname{Var}(G)}$, which, unlike $(1 - \operatorname{Var}(G - V)/\operatorname{Var}(G))_+$, stays positive when the critic explains none of the variance, as in early sparse-reward training. The two terms are combined by a geometric mean, so a low value of either lowers q_t, and q_t is smoothed with one EMA ($\beta = 0.9$).

C Improvement guarantees: proofs

Throughout, $\pi = \pi_{\theta_{\mathrm{old}}}$, $\pi' = \pi_\theta$, $\bar{A}(s) = \mathbb{E}_{a\sim\pi'}[A^\pi(s, a)]$, $\xi = \max_s |\bar{A}(s)|$, and all visitation measures are the $(1 - \gamma)$-normalized ones of Section 2.

Assumption 1 (Bounded ratios)**.** *There are constants $0 < w_{\min} \leq 1 \leq w_{\max} < \infty$ such that $w_\theta(s, a) \in [w_{\min}, w_{\max}]$ for $\nu_\rho^{\pi_{\theta_{\mathrm{old}}}}$-almost every (s, a).*

We need Assumption 1 to bound the total variation distance by D_α (Lemma C.3): on $[w_{\min}, w_{\max}]$ we have $f''_\alpha(w) = w^{\alpha-2} \geq m_\alpha = \min(w_{\min}^{\alpha-2}, w_{\max}^{\alpha-2})$. For $\alpha = 2$, $m_2 = 1$ and the assumption is not needed.

Assumption 2 (Stationary interval)**.** *$w_\theta(s, a) \in W_{\alpha,\epsilon} = [\, w^\star_-,\, w^\star_+\,]$ for $\nu_\rho^{\pi_{\theta_{\mathrm{old}}}}$-almost every (s, a), where $w^\star_+$ and $w^\star_-$ are the stationary ratios equation 8 for $\hat{A} > 0$ and $\hat{A} < 0$.*

Assumption 2 is Assumption 1 with $[w_{\min}, w_{\max}] = W_{\alpha,\epsilon}$. We use this interval because the objective equation 7 sets it: by Proposition 3.1, outside $W_{\alpha,\epsilon}$ every per-sample objective increases toward $W_{\alpha,\epsilon}$. Since samples share parameters (Section 3.1), the ratios can still leave $W_{\alpha,\epsilon}$ during an update, so we state this as an assumption.

Lemma C.1 (Performance difference; Kakade & Langford, 2002)**.** $J(\pi') - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi'}_\rho}[\bar{A}(s)]$.

Lemma C.2 (Visitation shift; Achiam et al., 2017)**.** $\big\|d^{\pi'}_\rho - d^{\pi}_\rho\big\|_1 \leq \frac{2\gamma}{1-\gamma}\,\mathbb{E}_{s\sim d^{\pi}_\rho}\big[D_{\mathrm{TV}}(\pi'(\cdot|s),\, \pi(\cdot|s))\big]$.

Lemma C.3 (Generalized Pinsker for the α-family)**.** *For distributions $P \ll Q$ on $\mathcal{A}$ with ratio $w = dP/dQ \in [w_{\min}, w_{\max}]$ and $m_\alpha = \min(w_{\min}^{\alpha-2}, w_{\max}^{\alpha-2})$: $D_{\mathrm{TV}}(P, Q)^2 \leq \frac{1}{2m_\alpha} D_\alpha(P\|Q)$.*

Proof. $f''_\alpha(w) = w^{\alpha-2}$ is monotone on $[w_{\min}, w_{\max}]$ (increasing iff $\alpha > 2$), so $f''_\alpha \geq m_\alpha$ there. Since $f_\alpha(1) = f'_\alpha(1) = 0$, the integral form of the remainder gives $f_\alpha(w) \geq \frac{m_\alpha}{2}(w-1)^2$ pointwise on the interval, hence $D_\alpha(P\|Q) = \mathbb{E}_Q[f_\alpha(w)] \geq \frac{m_\alpha}{2}\chi^2(P\|Q)$. By Cauchy–Schwarz, $\|P-Q\|_1^2 = (\mathbb{E}_Q|w-1|)^2 \leq \mathbb{E}_Q[(w-1)^2] = \chi^2(P\|Q)$, and $D_{\mathrm{TV}} = \frac{1}{2}\|P-Q\|_1$ gives $4D_{\mathrm{TV}}^2 \leq \chi^2 \leq \frac{2}{m_\alpha}D_\alpha$. This is an interval-restricted form of the f-divergence Pinsker inequalities of Gilardoni (2010) and Sason & Verdú (2016). For $\alpha = 2$, $m_2 = 1$ and the chain $4D_{\mathrm{TV}}^2 \leq \chi^2 = 2D_2$ needs no interval. □

Proof of Theorem 4.1(a). By Lemmas C.1–C.2 and Hölder,

$$\begin{aligned}\left|J(\pi') - J(\pi) - L_\pi(\pi')\right| &= \tfrac{1}{1-\gamma}\left|\mathbb{E}_{d_\rho^{\pi'}}[\bar{A}] - \mathbb{E}_{d_\rho^{\pi}}[\bar{A}]\right| \leq \tfrac{\xi}{1-\gamma}\|d_\rho^{\pi'} - d_\rho^{\pi}\|_1 \\ &\leq \tfrac{2\gamma\xi}{(1-\gamma)^2}\,\mathbb{E}_{d_\rho^{\pi}}[D_{\mathrm{TV}}] \;\leq\; \tfrac{2\gamma\xi}{(1-\gamma)^2}\sqrt{\mathbb{E}_{d_\rho^{\pi}}[D_{\mathrm{TV}}^2]} \;\leq\; \tfrac{2\gamma\xi}{(1-\gamma)^2}\sqrt{\tfrac{\bar{D}_\alpha}{2m_\alpha}}, \quad (25)\end{aligned}$$

using Jensen and the generalized Pinsker lemma state-wise. Rearranging gives equation 13. For the second claim, let $M_t(\pi') = J(\pi_t) + L_{\pi_t}(\pi') - C_\alpha\sqrt{\bar{D}_\alpha(\pi')}$. Then $J(\pi') \geq M_t(\pi')$ and $M_t(\pi_t) = J(\pi_t)$, since the surrogate and the penalty vanish at $\pi' = \pi_t$. Hence any π_{t+1} with $M_t(\pi_{t+1}) \geq M_t(\pi_t)$, in particular the maximizer of M_t, satisfies $J(\pi_{t+1}) \geq J(\pi_t)$. This is the argument of Schulman et al. (2015, Eq. 10), with the maximum KL penalty replaced by the average α-divergence.

The bound equation 13 also holds for any measurable state-dependent $\alpha(s)$, with m_α replaced by the harmonic mean $m_* = (\mathbb{E}_{s\sim d_\rho^\pi}[m_{\alpha(s)}^{-1}])^{-1}$. To see this, apply the Pinsker lemma with the state's own $\alpha(s)$, then Cauchy–Schwarz, which gives $\mathbb{E}_{d_\rho^\pi}[D_{\mathrm{TV}}] \leq \mathbb{E}\big[\sqrt{\frac{D_{\alpha(s)}}{2m_{\alpha(s)}}}\big] \leq \sqrt{\mathbb{E}[\frac{1}{2m_{\alpha(s)}}]\cdot\mathbb{E}[D_{\alpha(s)}]} = \frac{1}{\sqrt{2m_*}}\sqrt{\bar{D}_{\alpha(\cdot)}}$.

For $\alpha_t \in [\alpha_{\min}, \alpha_{\max}]$, equation 13 holds at every iteration with $\alpha = \alpha_t$, so the argument above gives $J(\pi_{t+1}) \geq J(\pi_t)$ at every iteration, and $J(\pi_t)$, which is bounded above by $R/(1-\gamma)$, converges. The argument uses no property of the rule that chooses α_t. □

Lemma C.4 (Penalized form)**.** *Let $\bar{\xi} = \max_{s,a}|A^{\pi_{\theta_{\mathrm{old}}}}(s,a)|$, and let $\bar{C}_\alpha$ be C_α with ξ replaced by $\bar{\xi}$. Under Assumption 1, for any $\eta > 0$, $J(\pi_\theta) \geq J(\pi_{\theta_{\mathrm{old}}}) + L_{\pi_{\theta_{\mathrm{old}}}}(\pi_\theta) - \frac{1}{2\eta}\bar{D}_\alpha - \frac{\eta}{2}\bar{C}_\alpha^2$. With $\eta = \epsilon/\bar{a}$, where $\bar{a} = \mathbb{E}|\hat{A}|$, the penalty is $\frac{\bar{a}}{2\epsilon}\bar{D}_\alpha$, of the same form as the penalty in equation 7, and the remaining term $\epsilon\bar{C}_\alpha^2/(2\bar{a})$ is $\mathcal{O}(\epsilon)$.*

Proof. Since $\xi \leq \bar{\xi}$, equation 13 holds with $\bar{C}_\alpha$ in place of C_α. By the AM–GM inequality, $\bar{C}_\alpha\sqrt{\bar{D}_\alpha} \leq \frac{1}{2\eta}\bar{D}_\alpha + \frac{\eta}{2}\bar{C}_\alpha^2$ for any $\eta > 0$, which gives the first claim, and $\eta = \epsilon/\bar{a}$ gives the second. The penalty $\frac{\bar{a}}{2\epsilon}\bar{D}_\alpha$ is that of equation 7 with $|\hat{A}|$ replaced by its mean $\bar{a}$ and a factor $\frac{1}{2}$; Appendix A.1 explains why equation 7 uses $|\hat{A}|$. Since $\bar{C}_\alpha$ does not depend on θ, maximizing $L_{\pi_{\theta_{\mathrm{old}}}}(\pi_\theta) - \frac{\bar{a}}{2\epsilon}\bar{D}_\alpha$ maximizes the lower bound. Hence the penalized update improves J up to an $\mathcal{O}(\epsilon)$ term, for every α. □

Special cases. At $\alpha = 1$, $m_1 = w_{\max}^{-1}$, which gives an average-KL version of the bound of Schulman et al. (2015). At $\alpha = 2$, $m_2 = 1$ without any interval, the best constant in the family. As $\alpha \to \infty$ with $w_{\min} < 1$, $m_\alpha \to 0$ and $C_\alpha \to \infty$, so the bound gives no guarantee in this limit.

C.1 Proof of Theorem 4.1(b)

Proof of Theorem 4.1(b). We bound D_{TV} directly. Per state, $D_{\mathrm{TV}} = \frac{1}{2}\mathbb{E}_{\pi_{\theta_{\mathrm{old}}}}|w-1| \leq \frac{1}{2}\sup_{W_{\alpha,\epsilon}}|w-1| = r_\alpha/2$. To identify r_α, note that $F(\beta) = w_+^\star + w_-^\star - 2$, with $\beta = \alpha - 1$, strictly decreases in β, since $\log w_\pm^\star$ are slopes of chords from 0 to β of the concave maps $x \mapsto \log(1 \pm x\epsilon)$ (Appendix A), and $F(1) = 0$. Hence $r_\alpha = w_+^\star - 1$ for $\alpha \leq 2$ and $r_\alpha = 1 - w_-^\star$ for $\alpha \geq 2$, and the expansion $w_s^\star = 1 + \epsilon s + \frac{2-\alpha}{2}\epsilon^2 + \mathcal{O}(\epsilon^3)$ gives equation 14's series.

Since $\mathbb{E}_{a\sim\pi_{\theta_{\mathrm{old}}}}[A^{\pi_{\theta_{\mathrm{old}}}}(s,a)] = 0$, we have $\bar{A}(s) = \mathbb{E}_{a\sim\pi_{\theta_{\mathrm{old}}}}\big[(w-1)A^{\pi_{\theta_{\mathrm{old}}}}(s,a)\big]$ and Hölder's inequality with $|w-1| \leq r_\alpha$ gives $|\bar{A}(s)| \leq r_\alpha\,\mathbb{E}_{a\sim\pi_{\theta_{\mathrm{old}}}}|A^{\pi_{\theta_{\mathrm{old}}}}(s,a)| \leq r_\alpha A_{\max}$ (cf. Schulman et al., 2015, Lemma 2).

Substituting both bounds into the proof of part (a), before Lemma C.3 is used,

$$\left|J(\pi') - J(\pi) - L_\pi(\pi')\right| \;\le\; \frac{2\gamma\,\xi}{(1-\gamma)^2}\,\mathbb{E}_{d_\rho^\pi}[D_{\mathrm{TV}}] \;\le\; \frac{2\gamma}{(1-\gamma)^2}\left(r_\alpha A_{\max}\right)\frac{r_\alpha}{2} \;=\; \frac{\gamma A_{\max}}{(1-\gamma)^2}\,r_\alpha^2, \tag{26}$$

which is the bound of Schulman et al. (2015, Theorem 1) with total variation distance $r_\alpha/2$.

Finally, equation 14 gives $r_\alpha^2 - r_2^2 = |\alpha-2|\,\epsilon^3 + \mathcal{O}(\epsilon^4) \le c_{\max}\epsilon^3 + \mathcal{O}(\epsilon^4)$ for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. □

D Proofs of the sample-complexity results of Section 4.2

D.1 Setting and standing assumptions

Rewards lie in $[-R, R]$ with $R > 0$, so $|A^\pi| \le A_\infty \le 2R/(1-\gamma)$, and the map $(\theta, s, a) \mapsto \pi_\theta(a|s)$ is Borel. Let π^* be an optimal policy, that is, a maximizer of $J(\pi)$ over all policies, and $J^* = J(\pi^*)$. We use the following general assumptions on the policy class (Liu et al., 2020).

Assumption 3 (Smooth score)**.** *For all θ, θ' and (s, a), $\|\nabla_\theta \log \pi_\theta(a|s)\| \le G$ and $\|\nabla_\theta \log \pi_\theta(a|s) - \nabla_\theta \log \pi_{\theta'}(a|s)\| \le B\,\|\theta - \theta'\|$.*

Assumption 4 (Fisher non-degeneracy and approximation error)**.** *For all θ, the Fisher matrix $F_\rho(\theta) = \mathbb{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}}\left[\nabla_\theta \log \pi_\theta(a|s)\,\nabla_\theta \log \pi_\theta(a|s)^\top\right]$ satisfies $F_\rho(\theta) \succeq \mu_F I$ with $\mu_F > 0$, and the transferred compatible-function-approximation error satisfies*

$$\mathbb{E}_{(s,a)\sim\nu_\rho^{\pi^*}}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)\,\omega^{*\top}\nabla_\theta \log \pi_\theta(a|s)\right)^2\right] \;\le\; \epsilon_{\mathrm{bias}}, \qquad \omega^* = F_\rho(\theta)^\dagger \nabla J(\theta). \tag{27}$$

This normalization of ϵ_{bias} follows Liu et al. (2020). In addition to these general assumptions, we assume that the Hessian of the log-policy is Lipschitz.

Assumption 5 (Lipschitz Hessian)**.** *For all θ, θ' and (s, a), $\|\nabla_\theta^2 \log \pi_\theta(a|s) - \nabla_\theta^2 \log \pi_{\theta'}(a|s)\| \le B_2\,\|\theta - \theta'\|$.*

This assumption is reasonable in our setting, since in the overparameterized regime the Hessian of the log-policy is smooth for networks with smooth activations, such as the $\tanh$ networks we use on MiniGrid and MuJoCo. Lemma D.2 uses it only for the continuity of $\nabla_\theta^2 \log \pi_\theta$, and it enters the bounds only through Lemma D.3.

Write $G_\omega = 2RG/((1-\gamma)^2\mu_F)$, so that $\|\omega^*\| \le \|\nabla J\|/\mu_F \le G_\omega/2$, and fix the trust constant $c_r = 2G_\omega$ of equation 15. On the trust ball $\|\theta' - \theta_k\| \le c_r\eta$ every ratio obeys $|\log w| \le G c_r \eta$, hence $|w-1| \le c_w\epsilon$ with $c_w = 2Gc_r(1-\gamma)/a_0$ for $\epsilon \le \epsilon_0$. Unbiased advantage estimates come from the geometric-stopping sampler of Agarwal et al. (2021, Algorithm 3), each costing $\mathcal{O}(1/(1-\gamma))$ expected environment steps, with the explicit scales of equation 89.

D.2 Outer-loop analysis

Lemma D.1 (Mondal & Aggarwal, 2024, Lemma 3)**.** *Let θ_0 be fixed, $\eta > 0$ and $K \ge 1$. For $k = 0, \dots, K-1$ let $\theta_{k+1} = \theta_k + \eta\,\omega_k$ with any random ω_k satisfying $\mathbb{E}\|\omega_k\|^2 < \infty$, and let $\omega_k^* = F_\rho(\theta_k)^\dagger \nabla J(\theta_k)$. Under Assumptions 3 and 4, if $\mathrm{KL}_0 = \mathbb{E}_{s\sim d_\rho^{\pi^*}}\left[D_{\mathrm{KL}}\big(\pi^*(\cdot|s)\,\|\,\pi_{\theta_0}(\cdot|s)\big)\right]$ is finite, then*

$$J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[J(\theta_k)\big] \;\le\; \frac{\sqrt{\epsilon_{\mathrm{bias}}}}{1-\gamma} + \frac{G}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\| + \frac{B\eta}{2K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k\|^2 + \frac{\mathrm{KL}_0}{\eta K}. \tag{28}$$

Lemma D.1 is Liu et al. (2020, Proposition 4.5) with the error $\|\omega_k - \omega_k^*\|$ in its first-order term replaced by the conditional bias $\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\|$.

Proof. Let $\mathrm{KL}_k = \mathbb{E}_{s\sim d_\rho^{\pi^*}}\big[D_{\mathrm{KL}}\big(\pi^*(\cdot|s)\,\|\,\pi_{\theta_k}(\cdot|s)\big)\big]$, $\nu^* = \nu_\rho^{\pi^*}$, $u_k = \nabla_\theta \log \pi_{\theta_k}(a|s)$ and $g_k = \mathbb{E}_{\nu^*}[u_k]$. For every k, we have

$$\begin{aligned}
\mathrm{KL}_k - \mathrm{KL}_{k+1} &= \mathbb{E}_{\nu^*}\big[\log \pi_{\theta_{k+1}}(a|s) - \log \pi_{\theta_k}(a|s)\big] \ge \eta\, g_k^\top \omega_k - \frac{B\eta^2}{2}\|\omega_k\|^2 \\
&= \eta\, \mathbb{E}_{\nu^*}\big[u_k^\top \omega_k^*\big] + \eta\, g_k^\top(\omega_k - \omega_k^*) - \frac{B\eta^2}{2}\|\omega_k\|^2 \\
&= \eta\big(J^* - J(\theta_k)\big) - \frac{\eta}{1-\gamma}\,\mathbb{E}_{\nu^*}\big[A^{\pi_{\theta_k}}(s,a) - (1-\gamma)\,u_k^\top \omega_k^*\big] \\
&\qquad + \eta\, g_k^\top(\omega_k - \omega_k^*) - \frac{B\eta^2}{2}\|\omega_k\|^2 \\
&\ge \eta\big(J^* - J(\theta_k)\big) - \frac{\eta\sqrt{\epsilon_{\mathrm{bias}}}}{1-\gamma} + \eta\, g_k^\top(\omega_k - \omega_k^*) - \frac{B\eta^2}{2}\|\omega_k\|^2, \qquad (29)
\end{aligned}$$

where we have applied the definition of D_{KL} in the first equality, with $\mathrm{KL}_k \le \mathrm{KL}_0 + G\eta\sum_{j<k}\|\omega_j\| < \infty$ by the G-bounded score, the B-Lipschitz score in the first inequality, Lemma C.1 with $\pi' = \pi^*$ and $\pi = \pi_{\theta_k}$ in the third equality, and Jensen's inequality with equation 27 at $\theta = \theta_k$ in the second inequality. We also have

$$\mathbb{E}\big[g_k^\top(\omega_k - \omega_k^*)\big] = \mathbb{E}\big[g_k^\top\big(\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big)\big] \ge -G\,\mathbb{E}\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\|, \qquad (30)$$

where we have applied the tower property in the equality, since g_k and ω_k^* are functions of θ_k, and $\|g_k\| \le G$ in the inequality. Taking expectations in equation 29 and applying equation 30, we obtain

$$J^* - \mathbb{E}\big[J(\theta_k)\big] \le \frac{\sqrt{\epsilon_{\mathrm{bias}}}}{1-\gamma} + G\,\mathbb{E}\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\| + \frac{B\eta}{2}\mathbb{E}\|\omega_k\|^2 + \frac{\mathbb{E}[\mathrm{KL}_k] - \mathbb{E}[\mathrm{KL}_{k+1}]}{\eta},$$

where the expectations are finite since $\mathbb{E}[\mathrm{KL}_k] \le \mathrm{KL}_0 + G\eta\sum_{j<k}\mathbb{E}\|\omega_j\|$, $\|\omega_k^*\| \le G_\omega/2$ and $\mathbb{E}\|\omega_k\|^2 < \infty$. Summing over $k = 0, \dots, K-1$, dividing by K and using $\mathrm{KL}_K \ge 0$ gives equation 28. □

In Proposition D.5, we bound the conditional bias of our update by the model bias $C_{\mathrm{prox}}(\alpha)\,\eta$ (Lemma D.2), the statistical bias C_N/N (Lemma D.3) and the inner-loop error $c_r e^{-c_0 T}$ (Lemma D.4), giving Theorem 4.2.

D.3 The proximal step is an inexact natural-gradient step

Lemma D.2 (Geometry equivalence and model bias)**.** *Assume $R > 0$, and let $\bar{L}_k(\theta') = \mathbb{E}_{\nu_\rho^{\pi_{\theta_k}}}\big[A^{\pi_{\theta_k}} w(\theta') - \frac{a_0}{\epsilon} f_\alpha(w(\theta'))\big]$ be the population surrogate. There are $\epsilon_{\mathrm{prox}}, c_1 > 0$, given in equation 32 and equation 33 below, such that for $\epsilon \le \epsilon_{\mathrm{prox}}$*

- *(a) $\nabla \bar{L}_k(\theta_k) = (1-\gamma)\nabla J(\theta_k)$, and on the trust ball $\bar{L}_k$ is μ-strongly concave with $\mu \ge a_0\mu_F/(2\epsilon)$, uniformly in $\alpha \in [\alpha_{\min}, \alpha_{\max}]$,*
- *(b) with $\eta = \epsilon(1-\gamma)/a_0$, $\bar{L}_k$ has a unique maximizer $\bar{\theta}_{k+1}$ on the trust ball, which satisfies $\|\bar{\theta}_{k+1} - \theta_k\| \le c_r\eta/3$, so that it is interior, and*

$$\Big\|\frac{\bar{\theta}_{k+1} - \theta_k}{\eta} - \omega_k^*\Big\| \;\le\; C_{\mathrm{prox}}(\alpha)\,\eta, \qquad C_{\mathrm{prox}}(\alpha) \;\le\; c_1 + \frac{G^3 G_\omega^2}{2\mu_F}\,|\alpha - 2|. \qquad (31)$$

The constants of the lemma are

$$\begin{aligned}
\epsilon_{\mathrm{prox}} &= \min\Big\{\frac{2\log(3/2)}{(2+c_{\max})\,c_w},\ \frac{G^2}{(G^2+B)\,c_w},\ \frac{a_0\mu_F}{4C_{\mathrm{dev}}}\Big\}, \\
C_{\mathrm{dev}} &= \frac{(1-\gamma)\,c_r G}{4}\big((18+9c_{\max})G^2 + 14B\big) + A_\infty(G^2+B),
\end{aligned} \qquad (32)$$

with $c_{\max} = \max(2-\alpha_{\min},\, \alpha_{\max}-2) \ge |\alpha - 2|$ as in Theorem 4.1, and

$$\begin{aligned}
c_1 &= \frac{(G^2+B)\,G_\omega}{\mu_F}\Big(G\,G_\omega + \frac{A_\infty}{1-\gamma}\Big) \qquad (33) \\
&\le \frac{2G(G^2+B)\,G_\omega^2}{\mu_F},
\end{aligned}$$

where the inequality uses $A_\infty \le 2R/(1-\gamma)$, $G_\omega = 2RG/((1-\gamma)^2\mu_F)$ and $\mu_F \le \|F_\rho(\theta_k)\| \le G^2$. In the proof k is fixed, and we write $\nu_k = \nu_\rho^{\pi_{\theta_k}}$, $F_k = F_\rho(\theta_k)$, $A = A^{\pi_{\theta_k}}(s,a)$, $w = w(\theta')$, $z = G\|\theta' - \theta_k\|$, $u_{\theta'} = \nabla_\theta \log \pi_{\theta'}(a|s)$, $U_{\theta'} = u_{\theta'}u_{\theta'}^\top + \nabla_\theta^2 \log \pi_{\theta'}(a|s)$ and $H(\theta') = -\nabla^2 \bar{L}_k(\theta')$. The bounded, B-Lipschitz score of Appendix D.1 gives $\|u_{\theta'}\| \le G$ and $\|U_{\theta'}\| \le G^2 + B$. The assumption $R > 0$ is kept to the end of this appendix; for $R = 0$ we have $J \equiv 0$ and equation 84 is trivial, and for $R > 0$ the constants G_ω, c_r and c_w are positive.

Proof. We bound the ratios and the curvature of $\bar{L}_k$ for (a), and then the step $\bar{\theta}_{k+1} - \theta_k$ and its distance to $\eta\,\omega_k^*$ for (b).

Bounding the ratios. As for the ratio envelope of Appendix D.1, the fundamental theorem of calculus along the segment from θ_k to θ' and the G-bounded score give, for every θ',

$$|\log w| \le z. \tag{34}$$

Therefore, for every θ' in the trust ball, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\epsilon \le \epsilon_{\text{prox}}$, we have

$$(2 + |\alpha - 2|)\, z \le \frac{(2 + c_{\max})\, c_w \epsilon}{2} \le \log \tfrac{3}{2}, \tag{35}$$

$$(G^2 + B)\, z \le \frac{(G^2 + B)\, c_w \epsilon}{2} \le \frac{G^2}{2}, \tag{36}$$

where we have applied $|\alpha - 2| \le c_{\max}$ and $z \le G c_r \eta = c_w \epsilon/2$ in the first inequalities, and the first and second entries of equation 32 in the second inequalities. Moreover, every v between 1 and w satisfies $|\log v| \le z$, so for every real c

$$v^c \le e^{|c|\, z}, \tag{37}$$

$$|v^c - 1| \le |c|\, z\, e^{|c|\, z}, \tag{38}$$

by $v^c = e^{c \log v}$ and $|e^y - 1| \le |y|\, e^{|y|}$. Therefore,

$$|f'_\alpha(w) - (w-1)| = \Big| \int_1^w \big(v^{\alpha-2} - 1\big)\, dv \Big| \le |\alpha - 2|\, z\, e^{|\alpha-2|z}\, |w - 1|, \tag{39}$$

where we have applied $f'_\alpha(1) = 0$ and $f''_\alpha(v) = v^{\alpha-2}$ in the equality, and equation 38 with $c = \alpha - 2$ in the inequality. In addition, the bounded, B-Lipschitz score gives

$$\big\|u_{\theta'}u_{\theta'}^\top - u_{\theta_k}u_{\theta_k}^\top\big\| \le 2G\, \|u_{\theta'} - u_{\theta_k}\| \le 2GB\, \|\theta' - \theta_k\|, \tag{40}$$

and, since $\pi_{\theta_k}(a|s) > 0$ for all (s,a),

$$\mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)}[w] = \int \pi_{\theta'}(a|s)\, da = 1. \tag{41}$$

Bounding the gradient and the Hessian. Next, with $\nabla w = w\, u_{\theta'}$ and $\nabla^2 w = w\, U_{\theta'}$, for every θ' with $\|\theta' - \theta_k\| < 2c_r\eta$ we have

$$\nabla \bar{L}_k(\theta') = \mathbb{E}_{\nu_k}\Big[\big(A - \tfrac{a_0}{\epsilon} f'_\alpha(w)\big)\, w\, u_{\theta'}\Big], \tag{42}$$

$$H(\theta') = \frac{a_0}{\epsilon}\, \mathbb{E}_{\nu_k}\Big[f''_\alpha(w)\, w^2\, u_{\theta'}u_{\theta'}^\top + f'_\alpha(w)\, w\, U_{\theta'}\Big] - \mathbb{E}_{\nu_k}\big[A\, w\, U_{\theta'}\big], \tag{43}$$

where differentiating under $\mathbb{E}_{\nu_k}$, and the continuity of both expressions in θ', follow by dominated convergence. There, equation 34 confines w to $[e^{-c_w\epsilon}, e^{c_w\epsilon}]$, where f_α, f'_α and f''_α are bounded, $|A| \le A_\infty$, and $\nabla_\theta^2 \log \pi_{\theta'}$ is continuous by the B_2-Lipschitz assumption. At $\theta' = \theta_k$ this gives

$$\nabla \bar{L}_k(\theta_k) = \mathbb{E}_{\nu_k}\big[A\, u_{\theta_k}\big] = (1-\gamma)\nabla J(\theta_k), \tag{44}$$

where we have applied equation 42 with $w = 1$ and $f'_\alpha(1) = 0$ in the first equality, and the policy gradient theorem $(1-\gamma)\nabla J(\theta_k) = \mathbb{E}_{\nu_k}[A\, u_{\theta_k}]$ (Sutton et al., 2000) in the second equality. This is the first claim of (a). Moreover, by Liu et al. (2020, Lemma B.1),

$$(1-\gamma)\|\nabla J(\theta_k)\| \le \frac{RG}{1-\gamma}. \tag{45}$$

Bounding the curvature deviation. Next, comparing equation 43 with $F_k = \mathbb{E}_{\nu_k}[u_{\theta_k} u_{\theta_k}^\top]$ and with the $\alpha = 2$ member, for which $f_2''(w) = 1$ and $f_2'(w) = w - 1$, we have

$$H(\theta') - \frac{a_0}{\epsilon} F_k = P_\alpha(\theta') + P_0(\theta'), \tag{46}$$

$$P_\alpha(\theta') = \frac{a_0}{\epsilon}\, \mathbb{E}_{\nu_k}\Big[\big(f_\alpha''(w) - 1\big)\, w^2\, u_{\theta'} u_{\theta'}^\top + \big(f_\alpha'(w) - (w-1)\big)\, w\, U_{\theta'}\Big],$$

$$P_0(\theta') = \frac{a_0}{\epsilon}\, \mathbb{E}_{\nu_k}\Big[(w^2 - 1)\, u_{\theta'} u_{\theta'}^\top + u_{\theta'} u_{\theta'}^\top - u_{\theta_k} u_{\theta_k}^\top + (w-1)\, w\, U_{\theta'}\Big] - \mathbb{E}_{\nu_k}\big[A\, w\, U_{\theta'}\big].$$

For θ' in the trust ball, the first part satisfies

$$\begin{aligned}
\|P_\alpha(\theta')\| &\le \frac{a_0}{\epsilon}\, \mathbb{E}_{\nu_k}\Big[|f_\alpha''(w) - 1|\, w^2 G^2 + |f_\alpha'(w) - (w-1)|\, w\, (G^2 + B)\Big] \\
&\le \frac{a_0}{\epsilon}\, |\alpha - 2|\, z\, e^{|\alpha-2|z}\Big(e^{2z} G^2 + z\, e^{2z}(G^2 + B)\Big) \\
&= \frac{a_0}{\epsilon}\, |\alpha - 2|\, G^3\, \|\theta' - \theta_k\|\, e^{(2+|\alpha-2|)z}\Big(1 + \frac{(G^2 + B)\, z}{G^2}\Big) \\
&\le \frac{9 a_0}{4\epsilon}\, |\alpha - 2|\, G^3\, \|\theta' - \theta_k\|,
\end{aligned}$$

where we have applied equation 37, equation 38 with $c \in \{1, 2, \alpha - 2\}$ and equation 39 in the second inequality, and equation 35 and equation 36 in the third inequality. The second part satisfies

$$\begin{aligned}
\|P_0(\theta')\| &\le \frac{a_0}{\epsilon}\, \mathbb{E}_{\nu_k}\Big[|w^2 - 1|\, G^2 + \big\|u_{\theta'} u_{\theta'}^\top - u_{\theta_k} u_{\theta_k}^\top\big\| + |w - 1|\, w\, (G^2 + B)\Big] \\
&\quad + A_\infty (G^2 + B)\, \mathbb{E}_{\nu_k}[w] \\
&\le \frac{a_0}{\epsilon}\Big(2z\, e^{2z} G^2 + 2GB\, \|\theta' - \theta_k\| + z\, e^{2z}(G^2 + B)\Big) + A_\infty (G^2 + B) \\
&\le \frac{a_0}{\epsilon}\Big(3G^3 + 2GB + \tfrac{3}{2}\, G(G^2 + B)\Big)\|\theta' - \theta_k\| + A_\infty (G^2 + B) \\
&= \frac{a_0 G}{2\epsilon}\, (9G^2 + 7B)\, \|\theta' - \theta_k\| + A_\infty (G^2 + B),
\end{aligned}$$

where we have applied $|A| \le A_\infty$ in the first inequality, equation 37 and equation 38 with $c \in \{1, 2\}$, equation 40 and equation 41 in the second inequality, and equation 35 in the third inequality. Combining the two bounds with equation 46 yields, for every θ' in the trust ball,

$$\Big\|H(\theta') - \frac{a_0}{\epsilon} F_k\Big\| \;\le\; \frac{a_0 G}{4\epsilon}\big(18G^2 + 14B + 9G^2|\alpha - 2|\big)\, \|\theta' - \theta_k\| + A_\infty (G^2 + B). \tag{47}$$

Bounding the curvature from below. Next, by $F_k \succeq \mu_F I$ (see Appendix D.1), for every θ' in the trust ball we have

$$H(\theta') \succeq \frac{a_0}{\epsilon} F_k - \Big\|H(\theta') - \frac{a_0}{\epsilon} F_k\Big\|\, I \succeq \Big(\frac{a_0 \mu_F}{\epsilon} - C_{\mathrm{dev}}\Big) I \succeq \frac{3 a_0 \mu_F}{4\epsilon}\, I, \tag{48}$$

where we have applied equation 47 with $\|\theta' - \theta_k\| \le c_r \eta$, $a_0 \eta / \epsilon = 1 - \gamma$ and $|\alpha - 2| \le c_{\max}$ in the second inequality, and the third entry of equation 32 in the third inequality. Hence $\bar{L}_k$ is μ-strongly concave on the trust ball with $\mu \ge 3 a_0 \mu_F / (4\epsilon)$ (Nesterov, 2018, Theorem 2.1.11), which completes (a).

Bounding the step. Next, $\bar{L}_k$ is continuous and strongly concave on the compact convex trust ball, so it has a unique maximizer $\bar{\theta}_{k+1}$ there (Nesterov, 2018, Theorem 2.2.10). Write $\Delta = \bar{\theta}_{k+1} - \theta_k$ and $H_{\mathrm{avg}} = \int_0^1 H(\theta_k + t\Delta)\, dt$, so that $H_{\mathrm{avg}} \succeq \frac{3 a_0 \mu_F}{4\epsilon} I$ by equation 48 and convexity of the trust ball. By the fundamental theorem of calculus,

$$H_{\mathrm{avg}}\, \Delta = \nabla \bar{L}_k(\theta_k) - \nabla \bar{L}_k(\bar{\theta}_{k+1}). \tag{49}$$

Therefore,

$$\begin{aligned}
\frac{3 a_0 \mu_F}{4\epsilon}\, \|\Delta\|^2 \le \Delta^\top H_{\mathrm{avg}}\, \Delta &= \big\langle \nabla \bar{L}_k(\theta_k),\, \Delta \big\rangle + \big\langle \nabla \bar{L}_k(\bar{\theta}_{k+1}),\, \theta_k - \bar{\theta}_{k+1} \big\rangle \\
&\le (1 - \gamma) \big\langle \nabla J(\theta_k),\, \Delta \big\rangle \le \frac{RG}{1 - \gamma}\, \|\Delta\|,
\end{aligned}$$

where we have applied equation 49 in the equality, equation 44 and the optimality condition $\langle \nabla \bar{L}_k(\bar{\theta}_{k+1}), \theta_k - \bar{\theta}_{k+1} \rangle \le 0$ (Nesterov, 2018, Theorem 2.2.9) in the second inequality, and the Cauchy–Schwarz inequality with equation 45 in the third inequality. Hence

$$\|\Delta\| \le \frac{4\epsilon R G}{3a_0(1-\gamma)\,\mu_F} = \frac{2}{3}\,G_\omega\,\eta, \tag{50}$$

where we have applied $\epsilon/a_0 = \eta/(1-\gamma)$ and the definition of G_ω in the equality. Since $c_r = 2G_\omega$, we get $\|\Delta\| \le c_r\eta/3$, so $\bar{\theta}_{k+1}$ is interior and $\nabla \bar{L}_k(\bar{\theta}_{k+1}) = 0$. Therefore,

$$H_{\mathrm{avg}}\,\Delta = \nabla \bar{L}_k(\theta_k) = (1-\gamma)\nabla J(\theta_k) = \frac{a_0\eta}{\epsilon}\,F_k\,\omega_k^*, \tag{51}$$

where we have applied equation 49 in the first equality, equation 44 in the second equality, and $\omega_k^* = F_k^{-1}\nabla J(\theta_k)$ with $a_0\eta/\epsilon = 1-\gamma$ in the third equality.

Bounding the model bias. Finally, by equation 51, we know that

$$\Delta - \eta\,\omega_k^* = \frac{\epsilon}{a_0}\,F_k^{-1}\Big(\frac{a_0}{\epsilon}F_k\Delta - H_{\mathrm{avg}}\Delta\Big) = \frac{\epsilon}{a_0}\,F_k^{-1}\int_0^1 \Big(\frac{a_0}{\epsilon}F_k - H(\theta_k + t\Delta)\Big)\,dt\;\Delta. \tag{52}$$

Therefore,

$$\begin{aligned}
\|\Delta - \eta\,\omega_k^*\| &\le \frac{\epsilon}{a_0\mu_F}\int_0^1 \Big\|H(\theta_k + t\Delta) - \frac{a_0}{\epsilon}F_k\Big\|\,dt\;\|\Delta\| \\
&\le \frac{\epsilon}{a_0\mu_F}\int_0^1 \Big(\frac{a_0 G\,t}{4\epsilon}\big(18G^2 + 14B + 9G^2|\alpha-2|\big)\|\Delta\| + A_\infty(G^2+B)\Big)dt\;\|\Delta\| \\
&= \frac{G}{8\mu_F}\big(18G^2 + 14B + 9G^2|\alpha-2|\big)\|\Delta\|^2 + \frac{\eta\,A_\infty(G^2+B)}{(1-\gamma)\,\mu_F}\,\|\Delta\| \\
&\le \eta^2\Big(\frac{G\,G_\omega^2}{18\mu_F}\big(18G^2 + 14B + 9G^2|\alpha-2|\big) + \frac{2\,G_\omega A_\infty(G^2+B)}{3(1-\gamma)\,\mu_F}\Big) \\
&\le \eta^2\Big(c_1 + \frac{G^3 G_\omega^2}{2\mu_F}\,|\alpha-2|\Big),
\end{aligned}$$

where we have applied equation 52 and $\|F_k^{-1}\| \le 1/\mu_F$ in the first inequality, equation 47 at $\theta' = \theta_k + t\Delta$ in the second inequality, $\epsilon/a_0 = \eta/(1-\gamma)$ in the equality, equation 50 in the third inequality, and $18G^2 + 14B \le 18(G^2+B)$ with equation 33 in the last inequality. Dividing by η gives equation 31, which completes (b). □

The proof uses the B_2-Lipschitz assumption only for the continuity of $\nabla_\theta^2 \log \pi_\theta$, so ϵ_{prox}, C_{dev} and c_1 are free of B_2. The $|\alpha-2|$ term of C_{prox} comes only from P_α, which is built from $f_\alpha''(w) - 1 = w^{\alpha-2} - 1$ with slope $f_\alpha'''(1) = \alpha - 2$ at $w = 1$ (see equation 9), while P_0 is the same for all members. The bound adds the two parts, but in the model bias they enter as vectors that can partly cancel, so we make no claim that $\alpha = 2$ minimizes the model bias.

D.4 Conditional bias of the batch proximal point

Throughout this subsection k is fixed, and $\Pr$ and $\mathbb{E}$ are conditional on θ_k, over the fresh batch, whose law given θ_k is a probability kernel; $(\theta, s, a) \mapsto \pi_\theta(a|s)$ and $(\theta_k, \theta') \mapsto \bar{L}_k(\theta')$ are Borel. Let d be the dimension of θ, $\mathcal{B}_k = \{\theta' : \|\theta' - \theta_k\| \le c_r\eta\}$ the trust ball, $\mathcal{B}_k^+ = \{\theta' : \|\theta' - \theta_k\| < 2c_r\eta\}$, $\bar{\theta} = \bar{\theta}_{k+1}$ the population maximizer of Lemma D.2, and $c_{\max}$ as in Theorem 4.1. For the i-th sample and $\theta' \in \mathbb{R}^d$ we write $w = w_i(\theta')$, $u = \nabla \log \pi_{\theta'}(a_i|s_i)$, $U = uu^\top + \nabla^2 \log \pi_{\theta'}(a_i|s_i)$, $\phi_1(w) = f_\alpha''(w)\,w^2$ and $\phi_2(w) = f_\alpha'(w)\,w$, and set

$$\ell_i(\theta') = \hat{A}_i\,w_i(\theta') - \frac{a_0}{\epsilon}f_\alpha\big(w_i(\theta')\big), \qquad \hat{L}_k = \frac{1}{N}\sum_{i=1}^N \ell_i,$$

$$H_i(\theta') = -\nabla^2\ell_i(\theta'), \qquad \hat{H}(\theta') = \frac{1}{N}\sum_{i=1}^N H_i(\theta'), \qquad H(\theta') = -\nabla^2\bar{L}_k(\theta'),$$

so that $\hat{L}_k$ is the batch objective of equation 15. The triples $(s_i, a_i, \hat{A}_i)$ are independent and identically distributed with $(s_i, a_i) \sim \nu_\rho^{\pi_{\theta_k}}$ and $\mathbb{E}[\hat{A}_i \mid s_i, a_i] = A^{\pi_{\theta_k}}(s_i, a_i)$. At $\bar{\theta}$ we drop the argument and write

$$g_i = \nabla \ell_i(\bar{\theta}), \qquad \hat{g} = \nabla \hat{L}_k(\bar{\theta}), \qquad H_i = H_i(\bar{\theta}), \qquad H = H(\bar{\theta}), \qquad \Delta_H = \hat{H}(\bar{\theta}) - H,$$

$$m_N = \frac{1}{N}\sum_{i=1}^{N} |\hat{A}_i|, \qquad \Delta_\theta = \hat{\theta}_{k+1} - \bar{\theta},$$

$$\hat{H}_{\mathrm{avg}} = \int_0^1 \hat{H}(\bar{\theta} + t\Delta_\theta)\, dt, \qquad \Delta_{\mathrm{avg}} = \hat{H}_{\mathrm{avg}} - \hat{H}(\bar{\theta}),$$

together with $M_N = C_3/\epsilon + M_w m_N$ and the dilation $\mathcal{D}(y) = \begin{pmatrix} 0 & y^\top \\ y & 0 \end{pmatrix}$ of $y \in \mathbb{R}^d$, which is linear in y with $\lambda_{\max}(\mathcal{D}(y)) = \|\mathcal{D}(y)\| = \|y\|$.

Imported inequalities. The sub-exponential norm is $\|X\|_{\psi_1} = \inf\{c > 0 : \mathbb{E}\exp(|X|/c) \le 2\}$, and for a random vector or matrix it is that of $\|X\|$. It is a monotone norm (Vershynin, 2018, Section 2.7) with $\|c\|_{\psi_1} = |c|/\log 2$ for constants, and the infimum is attained by monotone convergence, so $\mathbb{E}\exp(|X|/K) \le 2$ for every $K > 0$ with $K \ge \|X\|_{\psi_1}$. For every such K, Markov's inequality, $e^x \ge 1 + x^p/p!$ for $x \ge 0$, and Jensen's inequality give

$$\Pr(|X| > t) \le 2\exp\Big(-\frac{t}{K}\Big), \qquad \mathbb{E}|X|^p \le p!\, K^p, \qquad \mathbb{E}|X| \le K\log 2, \tag{53}$$

for all $t \ge 0$ and integers $p \ge 1$. The advantage estimates of Agarwal et al. (2021, Algorithm 3) satisfy $\|\hat{A}_i\|_{\psi_1} \le A_\psi$ and $\mathbb{E}[\hat{A}_i^2] \le A_2$ with $A_\psi = 4R/(1-\gamma)$ and $A_2 = 16R^2/(1-\gamma)^2$, by equation 89 in the proof of Theorem 4.2, which uses only the sampler and $|r| \le R$; this subsection uses only $A_\psi > 0$ and $A_2 < \infty$. By the matrix Bernstein inequality for sub-exponential summands (Tropp, 2012, Theorem 6.2), if $X_1, \dots, X_N$ are independent random symmetric $n \times n$ matrices with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^p \preceq \frac{p!}{2}\varsigma^{p-2}\Gamma_i^2$ for all integers $p \ge 2$, where $\varsigma > 0$ and the Γ_i are deterministic symmetric matrices, then with $\sigma_B^2 = \|\sum_i \Gamma_i^2\|$ and for all $t \ge 0$,

$$\Pr\Big(\lambda_{\max}\Big(\sum_{i=1}^{N} X_i\Big) \ge t\Big) \;\le\; n\exp\Big(-\frac{t^2/2}{\sigma_B^2 + \varsigma t}\Big). \tag{54}$$

Finally, let $\mathcal{K} \subseteq \mathbb{R}^d$ be closed and convex with nonempty interior, and let q be twice continuously differentiable on an open set containing $\mathcal{K}$ with $-\nabla^2 q(x) \succeq \mu_q I$ for all $x \in \mathcal{K}$. Then q is μ_q-strongly concave on $\mathcal{K}$, and for all $x, y \in \mathcal{K}$ and every maximizer $x^\star$ of q on $\mathcal{K}$ (Nesterov, 2018, Theorems 2.1.9, 2.1.11 and 2.2.9, Corollary 2.2.1)

$$\begin{aligned} &\big\langle \nabla q(x) - \nabla q(y),\, y - x\big\rangle \ge \mu_q \|y - x\|^2, \qquad \big\langle \nabla q(x^\star),\, y - x^\star\big\rangle \le 0, \\ &q(x^\star) - q(y) \ge \frac{\mu_q}{2}\|y - x^\star\|^2. \end{aligned} \tag{55}$$

Constants. With c_w and c_r of Appendix D.1 and the advantage bounds A_ψ and A_2, set

$$\begin{aligned} &M_w = 4\,(G^3 + GB + B_2), \qquad C_3 = 18\,a_0\,(1 + |\alpha - 2|)\,(G^3 + GB + B_2), \\ &\bar{C}_3 = 18\,a_0\,(1 + c_{\max})\,(G^3 + GB + B_2), \qquad c_+ = 6 \cdot 2^{c_{\max}}\,(1 + G + G^2 + B), \\ &\kappa_1 = 8a_0G^2, \qquad \kappa_2 = 2(G^2 + B), \qquad G_\ell = 2G\,(A_\psi + 3a_0c_w), \\ &G_{\ell,2} = 8G^2\,(A_2 + 4a_0^2c_w^2), \qquad \mu_\epsilon = \frac{a_0\mu_F}{2\epsilon}, \qquad b_H = \frac{4\kappa_1}{\epsilon}, \qquad t_b = \frac{\mu_F(1-\gamma)\,c_r}{16}, \\ &c_H = \frac{\mu_F^2}{327680\,G^4}, \qquad c_b = \frac{1}{6}\min\Big(\frac{t_b^2}{G_\ell^2}, \frac{t_b}{G_\ell}\Big), \qquad c_\star = \min\Big(c_H, c_b, \frac{1}{20}\Big), \\ &\Lambda_d = 1 + \log(2d), \qquad N_0 = \max\Big(\Big\lceil \frac{2}{c_\star}\log\frac{8d}{c_\star}\Big\rceil,\ \lceil \Lambda_d^2 \rceil\Big), \end{aligned} \tag{56}$$

so that μ_ϵ is the curvature lower bound of Lemma D.2(a). The threshold ϵ_0 of Proposition D.5 is

$$\begin{aligned} \epsilon_0 = \min\Big\{&\epsilon_{\mathrm{prox}},\ 1,\ \frac{1}{2c_w},\ \frac{2\log 2}{c_{\max}c_w},\ \frac{G^2}{(G^2+B)\,c_w}, \\ &\frac{\kappa_1}{3\kappa_2 A_\psi},\ \frac{a_0^2\mu_F}{16\,c_r(1-\gamma)\,(2\bar{C}_3 + 3M_wA_\psi)}\Big\}, \end{aligned} \tag{57}$$

with ϵ_{prox} of equation 32 and the fourth entry $+\infty$ when $c_{\max} = 0$; the third to fifth entries are implied by the first and are listed for reference. The good event and its failure budget are

$$\mathcal{E} = \underbrace{\Big\{\|\Delta_H\| \le \frac{\mu_\epsilon}{4},\ m_N \le 2A_\psi\Big\}}_{\text{curvature and mean advantage}} \cap \underbrace{\big\{\|\hat{g}\| \le t_b\big\}}_{\text{interiority}}, \tag{58}$$

$$p_N = 2d\, e^{-c_H N} + e^{-N/20} + (d+1)\, e^{-c_b N}. \tag{59}$$

Lemma D.3 (M-estimator conditional bias)**.** *Let $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\epsilon \le \epsilon_0$ with ϵ_0 as in equation 57, let $\mathcal{E}$ and p_N be as in equation 58 and equation 59, and let $\hat{\theta}_{k+1}$ be a point of the trust ball that is a Borel function of θ_k and the batch and maximizes the batch objective of equation 15 on the trust ball whenever $\mathcal{E}$ holds. Then $\Pr(\mathcal{E}^c \mid \theta_k) \le p_N$, and $p_N \le 1/N$ for $N \ge N_0$. On $\mathcal{E}$, for every θ' in the trust ball,*

$$\frac{\mu_\epsilon}{2} I \ \preceq\ -\nabla^2 \hat{L}_k(\theta') \ \preceq\ \Big(\frac{\kappa_1}{\epsilon} + 2\kappa_2 A_\psi\Big) I, \tag{60}$$

so the batch objective is $\frac{\mu_\epsilon}{2}$-strongly concave on the trust ball and its maximizer $\hat{\theta}_{k+1}$ there is unique and satisfies $\|\hat{\theta}_{k+1} - \theta_k\| \le \frac{7}{12} c_r \eta$. An admissible $\hat{\theta}_{k+1}$ exists. Moreover, for $N \ge N_0$, conditionally on θ_k,

$$\big\|\mathbb{E}[\hat{\theta}_{k+1} \mid \theta_k] - \bar{\theta}_{k+1}\big\| \ \le\ \frac{C_b\, \epsilon}{N}, \qquad \mathrm{Var}\big(\hat{\omega}_k \mid \theta_k\big) \ \le\ \frac{\sigma_\omega^2}{N}, \tag{61}$$

where $\hat{\omega}_k = (\hat{\theta}_{k+1} - \theta_k)/\eta$, $\mathrm{Var}(X \mid \theta_k) = \mathbb{E}\big[\|X - \mathbb{E}[X \mid \theta_k]\|^2 \mid \theta_k\big]$, and the constants C_b and σ_ω^2, given in equation 80 and equation 81, are free of ϵ and N. In direction units, $\big\|\mathbb{E}[\hat{\omega}_k \mid \theta_k] - (\bar{\theta}_{k+1} - \theta_k)/\eta\big\| \le C_N/N$ with $C_N = C_b\, a_0/(1-\gamma)$.

Proof. We bound the per-sample derivatives, the moments and $\Pr(\mathcal{E}^c)$, locate the batch maximizer on $\mathcal{E}$, expand it around $\bar{\theta}$, and bound each term of the expansion.

Bounding the per-sample derivatives. By equation 34, every $\theta' \in \mathcal{B}_k$ has $|\log w| \le G c_r \eta = c_w \epsilon/2 \le \frac{1}{4}$, the last by the third entry of equation 57. For every w with $|\log w| \le c_w\epsilon/2$ we have

$$\tfrac{3}{4} \le w \le \tfrac{4}{3}, \quad |w-1| \le c_w\epsilon, \quad \tfrac{1}{2} \le f''_\alpha(w) \le 2, \quad |f'_\alpha(w)| \le 2c_w\epsilon, \quad |f'''_\alpha(w)| \le \tfrac{8}{3}|\alpha-2|, \tag{62}$$

where we have applied $e^{-1/4} \ge \frac{3}{4}$ and $e^{1/4} \le \frac{4}{3}$ in the first bound, $|e^x - 1| \le 2|x|$ for $|x| \le \frac{1}{4}$ in the second bound, $f''_\alpha(w) = w^{\alpha-2}$ with $|(\alpha-2)\log w| \le c_{\max} c_w \epsilon/2 \le \log 2$ by the fourth entry of equation 57 in the third bound, the mean value theorem with $f'_\alpha(1) = 0$ and the second and third bounds in the fourth bound, and $f'''_\alpha(w) = (\alpha-2) f''_\alpha(w)/w$ in the fifth bound. Combining this with $\phi'_1 = f'''_\alpha w^2 + 2 f''_\alpha w$ and $\phi'_2 = f''_\alpha w + f'_\alpha$ yields, for the same w,

$$|\phi_1(w)| \le 4, \qquad |\phi_2(w)| \le \tfrac{8}{3}\, c_w\epsilon, \qquad |\phi'_1(w)| \le 5|\alpha-2| + 6, \qquad |\phi'_2(w)| \le 4, \tag{63}$$

where the last bound uses $c_w\epsilon \le \frac{1}{2}$. Next, since $\nabla w = w\, u$ and $\nabla^2 w = w\, U$, for every $\theta' \in \mathbb{R}^d$ we have

$$\nabla \ell_i(\theta') = \Big(\hat{A}_i - \frac{a_0}{\epsilon} f'_\alpha(w)\Big)\, w\, u, \qquad H_i(\theta') = \frac{a_0}{\epsilon}\big(\phi_1(w)\, uu^\top + \phi_2(w)\, U\big) - \hat{A}_i\, w\, U.$$

Therefore, for every $\theta' \in \mathcal{B}_k$,

$$\begin{aligned}
\|\nabla \ell_i(\theta')\| &\le \Big(|\hat{A}_i| + \frac{a_0}{\epsilon}|f'_\alpha(w)|\Big)\, w\, G \le \tfrac{4}{3}\, G\, \big(|\hat{A}_i| + 2a_0 c_w\big) \le \zeta_i, \\
\|H_i(\theta')\| &\le \frac{a_0}{\epsilon}\big(|\phi_1(w)|\, G^2 + |\phi_2(w)|\, (G^2 + B)\big) + |\hat{A}_i|\, w\, (G^2 + B) \\
&\le \frac{4a_0 G^2}{\epsilon} + \tfrac{8}{3}\, a_0 c_w (G^2 + B) + \tfrac{4}{3}(G^2 + B)\, |\hat{A}_i| \le \frac{\kappa_1}{\epsilon} + \kappa_2 |\hat{A}_i|,
\end{aligned}$$

where $\zeta_i = 2G(|\hat{A}_i| + 2a_0 c_w)$, and we have applied $\|u\| \le G$ and $\|U\| \le G^2 + B$ in the first inequalities, equation 62 and equation 63 in the second inequalities, and $\frac{4}{3} \le 2$ and the fifth entry of equation 57 in the last inequalities. On the other hand, for $\theta', \theta'' \in \mathcal{B}_k$ with the corresponding (w', u', U') and (w'', u'', U''), equation 62, the bounded, B-Lipschitz score and the B_2-Lipschitz Hessian of $\log \pi$ give

$$\begin{aligned}
&|w' - w''| \le \tfrac{4}{3}\, G\, \|\theta' - \theta''\|, \qquad \|u'u'^\top - u''u''^\top\| \le 2GB\, \|\theta' - \theta''\|, \\
&\|U' - U''\| \le (2GB + B_2)\, \|\theta' - \theta''\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}\|H_i(\theta') - H_i(\theta'')\| &\le \frac{a_0}{\epsilon}\Big(\sup|\phi_1'|\,|w'-w''|\,G^2 + \sup|\phi_1|\,\|u'u'^\top - u''u''^\top\|\Big)\\ &\quad + \frac{a_0}{\epsilon}\Big(\sup|\phi_2'|\,|w'-w''|\,(G^2+B) + \sup|\phi_2|\,\|U'-U''\|\Big)\\ &\quad + |\hat{A}_i|\Big(|w'-w''|\,(G^2+B) + \tfrac{4}{3}\,\|U'-U''\|\Big)\\ &\le \frac{a_0}{\epsilon}\Big(\big(\tfrac{20}{3}|\alpha-2| + \tfrac{40}{3}\big)G^3 + 16\,GB + \tfrac{4}{3}\,B_2\Big)\|\theta'-\theta''\|\\ &\quad + \tfrac{4}{3}\,|\hat{A}_i|\,\big(G^3 + 3GB + B_2\big)\,\|\theta'-\theta''\| \le \Big(\frac{C_3}{\epsilon} + M_w|\hat{A}_i|\Big)\|\theta'-\theta''\|,\end{aligned}$$

where the suprema are over $|\log w| \le c_w\epsilon/2$, and we have applied the triangle inequality and the mean value theorem for ϕ_1 and ϕ_2 in the first inequality, the previous display, equation 63 and $\frac{8}{3}c_w\epsilon \le \frac{4}{3}$ in the second inequality, and equation 56 in the third inequality. In addition, for every $\theta' \in \mathcal{B}_k^+$ we have

$$\begin{aligned}|\ell_i(\theta')| + \|\nabla\ell_i(\theta')\| + \|\nabla^2\ell_i(\theta')\| &\le |\hat{A}_i|\,w\,(1+G+G^2+B) + \frac{a_0}{\epsilon}\big(f_\alpha(w) + f_\alpha''(w)\,w^2G^2\big)\\ &\quad + \frac{a_0}{\epsilon}\,|f_\alpha'(w)|\,w\,(G+G^2+B)\\ &\le 2|\hat{A}_i|\,(1+G+G^2+B) + 2^{c_{\max}}\,(1+2G+6G^2+2B)\,\frac{a_0}{\epsilon}\\ &\le c_+\Big(|\hat{A}_i| + \frac{a_0}{\epsilon}\Big),\end{aligned}$$

where we have applied the triangle inequality in the first inequality, and equation 34 with $|\log w| \le 2Gc_r\eta = c_w\epsilon \le \frac{1}{2}$, which gives $w \le 2$, $f_\alpha(w) \le 2^{c_{\max}}/2$, $f_\alpha''(w) \le 2^{c_{\max}}$ and $|f_\alpha'(w)| \le 2^{c_{\max}}$, in the second inequality. Combining the bounds of this block yields, for $\theta', \theta'' \in \mathcal{B}_k$,

$$\begin{aligned}&\|\nabla\ell_i(\theta')\| \le \zeta_i, \qquad \|H_i(\theta')\| \le \frac{\kappa_1}{\epsilon} + \kappa_2|\hat{A}_i|,\\ &\|\hat{H}(\theta') - \hat{H}(\theta'')\| \le \Big(\frac{C_3}{\epsilon} + M_w m_N\Big)\|\theta'-\theta''\|,\\ &\|H(\theta') - H(\theta'')\| \le \Big(\frac{C_3}{\epsilon} + M_w A_\psi\Big)\|\theta'-\theta''\|,\end{aligned} \tag{64}$$

where the fourth bound follows from $H(\theta') = \mathbb{E}H_i(\theta')$, Jensen's inequality and $\mathbb{E}|\hat{A}_i| \le A_\psi \log 2$ from equation 53. Here $\bar{L}_k = \mathbb{E}\ell_i$ by the tower property, and dominated convergence with the previous display makes $\bar{L}_k$ twice continuously differentiable on $\mathcal{B}_k^+$ with $\nabla\bar{L}_k = \mathbb{E}\nabla\ell_i$ and $H = \mathbb{E}H_i$.

Bounding the population maximizer. First, Lemma D.2(b) applies since $\epsilon \le \epsilon_0 \le \epsilon_{\text{prox}}$, and gives

$$\|\bar{\theta} - \theta_k\| \;\le\; \frac{c_r\eta}{3}. \tag{65}$$

Therefore, since $\bar{\theta}$ is interior to $\mathcal{B}_k$, we have

$$\mathbb{E}g_i = 0, \qquad \mathbb{E}H_i = H, \qquad H(\theta') \succeq \mu_\epsilon I \;\text{ for all } \theta' \in \mathcal{B}_k, \qquad \|H^{-1}\| \le \frac{1}{\mu_\epsilon}, \tag{66}$$

where we have applied $\nabla\bar{L}_k(\bar{\theta}) = 0$ in the first identity, and equation 48 with $3a_0\mu_F/(4\epsilon) \ge \mu_\epsilon$ in the third and fourth bounds.

Bounding the moments. First, by the triangle inequality for $\|\cdot\|_{\psi_1}$, we know that

$$\begin{aligned}\|\zeta_i\|_{\psi_1} &\le 2G\,\|\hat{A}_i\|_{\psi_1} + \frac{4Ga_0c_w}{\log 2} \le G_\ell,\\ \|H_i - H\|_{\psi_1} &\le \Big\|\frac{\kappa_1}{\epsilon} + \kappa_2|\hat{A}_i| + \mathbb{E}\|H_i\|\Big\|_{\psi_1} \le \Big\|\frac{2\kappa_1}{\epsilon} + \kappa_2 A_\psi \log 2\Big\|_{\psi_1} + \kappa_2\,\|\hat{A}_i\|_{\psi_1}\\ &\le \frac{2\kappa_1}{\epsilon\log 2} + 2\kappa_2 A_\psi \le b_H,\end{aligned}$$

where we have applied $\|c\|_{\psi_1} = |c|/\log 2$ in the first inequality of the first chain and the third inequality of the second chain, $\|\hat{A}_i\|_{\psi_1} \le A_\psi$ in the second inequality of the first chain and the third inequality of the second chain, $2/\log 2 \le 3$ in the second inequality of the first chain, monotonicity and equation 64 with $\|H\| \le \mathbb{E}\|H_i\|$ in the first inequality of the second chain, equation 64 with $\mathbb{E}|\hat{A}_i| \le A_\psi \log 2$ in its second inequality, and $2/\log 2 \le 3$ with the sixth entry of equation 57 in its fourth inequality. Therefore, for $p \in \{2, 4\}$ we have

$$\big(\mathbb{E}\|H_i\|^p\big)^{1/p} \le \frac{\kappa_1}{\epsilon} + \kappa_2\big(\mathbb{E}|\hat{A}_i|^p\big)^{1/p} \le \frac{\kappa_1}{\epsilon} + 3\kappa_2 A_\psi \le \frac{2\kappa_1}{\epsilon},$$

where we have applied Minkowski's inequality in the first inequality, equation 53 with $(p!)^{1/p} \le 3$ in the second inequality, and the sixth entry of equation 57 in the third inequality. In addition, $\mathbb{E}\|g_i\|^2 \le \mathbb{E}\zeta_i^2 \le 8G^2(A_2 + 4a_0^2 c_w^2) = G_{\ell,2}$. Since the g_i are independent with $\mathbb{E}g_i = 0$, expanding $\|\sum_i g_i\|^4$ and dropping the terms in which an index appears once gives the standard bound

$$\mathbb{E}\|\hat{g}\|^4 \le \frac{N\,\mathbb{E}\|g_1\|^4 + 3N^2\big(\mathbb{E}\|g_1\|^2\big)^2}{N^4} \le \frac{36\,G_\ell^4}{N^2},$$

where the second inequality uses $\mathbb{E}\|g_1\|^p \le \mathbb{E}\zeta_1^p \le p!\,G_\ell^p$ from equation 53. Moreover,

$$\big(\mathbb{E}\|\Delta_H\|^4\big)^{1/4} \le \Big(\mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N \|H_i\|\Big)^4\Big)^{1/4} + \|H\| \le 2\big(\mathbb{E}\|H_1\|^4\big)^{1/4} \le \frac{4\kappa_1}{\epsilon},$$

by the triangle and Minkowski inequalities, $\|H\| \le (\mathbb{E}\|H_1\|^4)^{1/4}$ and the previous bound on $(\mathbb{E}\|H_i\|^4)^{1/4}$. Collecting these bounds with $\mathbb{E}\|\hat{g}\|^2 = \mathbb{E}\|g_1\|^2/N$ yields

$$\begin{aligned} &\mathbb{E}\|H_1\|^2 \le \frac{4\kappa_1^2}{\epsilon^2}, \qquad \mathbb{E}\|g_1\|^2 \le G_{\ell,2}, \qquad \mathbb{E}\|\hat{g}\|^2 \le \frac{G_{\ell,2}}{N}, \\ &\big(\mathbb{E}\|\hat{g}\|^4\big)^{1/4} \le \frac{\sqrt{6}\,G_\ell}{\sqrt{N}}, \qquad \big(\mathbb{E}\|\Delta_H\|^4\big)^{1/4} \le \frac{4\kappa_1}{\epsilon}. \end{aligned} \tag{67}$$

Bounding $\Pr(\mathcal{E}^c)$**.** First, let $X_1, \dots, X_N$ be independent random symmetric $n \times n$ matrices with $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_1} \le b$ for some $b > 0$. Then for all integers $p \ge 2$,

$$\mathbb{E}\Big(\frac{X_i}{N}\Big)^p \preceq \frac{\mathbb{E}\|X_i\|^p}{N^p}\, I \preceq \frac{p!}{2}\Big(\frac{b}{N}\Big)^{p-2}\frac{2b^2}{N^2}\, I,$$

by $Y^p \preceq \|Y\|^p I$ and equation 53. Therefore equation 54 with $\varsigma = b/N$ and $\Gamma_i = \sqrt{2}\,b\,I/N$, so that $\sigma_B^2 = 2b^2/N$, yields

$$\Pr\Big(\lambda_{\max}\Big(\frac{1}{N}\sum_{i=1}^N X_i\Big) \ge t\Big) \;\le\; n\exp\Big(-\frac{Nt^2}{2(2b^2 + bt)}\Big) \qquad \text{for all } t \ge 0. \tag{68}$$

Applying equation 68 to $\pm(H_i - H)$, which are independent and centered by equation 66, with $n = d$ and $b = b_H$ yields

$$\Pr\big(\|\Delta_H\| \ge t\big) \;\le\; 2d\exp\Big(-\frac{Nt^2}{2(2b_H^2 + b_H t)}\Big) \qquad \text{for all } t \ge 0. \tag{69}$$

Therefore, with $x_H = \mu_\epsilon/(4b_H)$, we have

$$\Pr\Big(\|\Delta_H\| > \frac{\mu_\epsilon}{4}\Big) \le 2d\exp\Big(-\frac{Nx_H^2}{2(2 + x_H)}\Big) \le 2d\exp\Big(-\frac{Nx_H^2}{5}\Big) = 2d\,e^{-c_H N},$$

where we have applied equation 69 at $t = x_H b_H$ in the first inequality, and $x_H = \mu_F/(256\,G^2) \le 1/256$ by $\mu_F \le G^2$ in the second inequality. On the other hand, $Y_i = |\hat{A}_i| - \mathbb{E}|\hat{A}_i|$ are independent and centered with $\|Y_i\|_{\psi_1} \le A_\psi + \mathbb{E}|\hat{A}_i|/\log 2 \le 2A_\psi$, so

$$\Pr\big(m_N > 2A_\psi\big) \le \Pr\Big(\frac{1}{N}\sum_{i=1}^N Y_i \ge A_\psi\Big) \le \exp\Big(-\frac{NA_\psi^2}{2(8A_\psi^2 + 2A_\psi^2)}\Big) = e^{-N/20},$$

where we have applied $\mathbb{E}|\hat{A}_i| \le A_\psi$ in the first inequality, and equation 68 with $n = 1$, $b = 2A_\psi$ and $t = A_\psi$ in the second inequality. Next, the dilations $\mathcal{D}(g_i)$ are independent and centered with $\|\mathcal{D}(g_i)\|_{\psi_1} \le \|\zeta_i\|_{\psi_1} \le G_\ell$, so with $x_b = t_b/G_\ell$,

$$\begin{aligned}\Pr\big(\|\hat{g}\| > t_b\big) &\le (d+1)\exp\Big(-\frac{N x_b^2}{2(2+x_b)}\Big) \le (d+1)\exp\Big(-\frac{N}{6}\min(x_b^2, x_b)\Big)\\ &= (d+1)\,e^{-c_b N},\end{aligned}$$

where we have applied equation 68 with $n = d+1$, $b = G_\ell$ and $\lambda_{\max}\big(\frac{1}{N}\sum_i \mathcal{D}(g_i)\big) = \|\hat{g}\|$ in the first inequality, and $x^2/(2(2+x)) \ge \min(x^2, x)/6$ in the second inequality. The union bound gives $\Pr(\mathcal{E}^c) \le p_N$. Finally, with $\beta_\star = \log(8d/c_\star)$, for $N \ge N_0$ we have

$$c_\star N - \log N \ge 2\beta_\star - \log\frac{2\beta_\star}{c_\star} = \log(8d) + \beta_\star - \log(2\beta_\star) \ge \log(8d),$$

where we have applied $N \ge 2\beta_\star/c_\star \ge 1/c_\star$, which holds since $\beta_\star \ge \log 160$, and the monotonicity of $N \mapsto c_\star N - \log N$ on $N \ge 1/c_\star$ in the first inequality, $\log(1/c_\star) = \beta_\star - \log(8d)$ in the equality, and $e^x \ge 2x$ in the second inequality. Therefore $p_N \le 5d\,e^{-c_\star N} \le 5/(8N) \le 1/N$, by $c_\star \le \min(c_H, c_b, 1/20)$, $3d + 2 \le 5d$ and the previous display.

Bounding the batch maximizer on $\mathcal{E}$. First, on $\mathcal{E}$ for every $\theta' \in \mathcal{B}_k$ we know that

$$\begin{aligned}\|\hat{H}(\theta') - H(\theta')\| &\le \|\hat{H}(\theta') - \hat{H}(\bar{\theta})\| + \|\Delta_H\| + \|H - H(\theta')\|\\ &\le 2c_r\eta\Big(\frac{2C_3}{\epsilon} + M_w(m_N + A_\psi)\Big) + \frac{\mu_\epsilon}{4}\\ &\le \frac{2c_r(1-\gamma)}{a_0}\big(2C_3 + 3M_w A_\psi\,\epsilon\big) + \frac{\mu_\epsilon}{4} \le \frac{\mu_\epsilon}{2},\end{aligned}$$

where we have applied equation 64 with $\|\theta' - \bar{\theta}\| \le 2c_r\eta$ in the second inequality, $m_N \le 2A_\psi$ and $\eta = \epsilon(1-\gamma)/a_0$ in the third inequality, and $C_3 \le \bar{C}_3$, $\epsilon \le 1$ and the last entry of equation 57 in the fourth inequality. Therefore, on $\mathcal{E}$ for every $\theta' \in \mathcal{B}_k$ we have

$$\begin{aligned}&\hat{H}(\theta') \succeq H(\theta') - \|\hat{H}(\theta') - H(\theta')\|\, I \succeq \frac{\mu_\epsilon}{2}\, I,\\ &\hat{H}(\theta') \preceq \frac{1}{N}\sum_{i=1}^N \|H_i(\theta')\|\, I \preceq \Big(\frac{\kappa_1}{\epsilon} + \kappa_2 m_N\Big) I \preceq \Big(\frac{\kappa_1}{\epsilon} + 2\kappa_2 A_\psi\Big) I,\end{aligned}$$

by equation 66, the previous display, equation 64 and $m_N \le 2A_\psi$. This proves equation 60, so on $\mathcal{E}$ the batch objective $\hat{L}_k$ has a unique maximizer $\hat{\theta}_{k+1}$ on $\mathcal{B}_k$ (Nesterov, 2018, Theorem 2.2.10). For measurability, let $(q_j)_{j\ge1}$ enumerate the rational points of the unit ball. By the third inequality of equation 55 and continuity, the maximizers of a strongly concave function over $\{\theta_k + c_r\eta\, q_j\}_{j\le n}$, with ties broken by the smallest index, converge to its maximizer on $\mathcal{B}_k$. Applied to $\bar{L}_k$, this makes $\bar{\theta}$ a Borel function of θ_k, hence, derivatives in θ' being pointwise limits of difference quotients, $\hat{g}$, Δ_H and $\mathcal{E}$ are measurable; applied to $\hat{L}_k$ on $\mathcal{E}$, with $\hat{\theta}_{k+1} = \theta_k$ off $\mathcal{E}$, it gives an admissible $\hat{\theta}_{k+1}$. Next, on $\mathcal{E}$ we know that

$$\frac{\mu_\epsilon}{2}\,\|\Delta_\theta\|^2 \le \big\langle \nabla\hat{L}_k(\bar{\theta}) - \nabla\hat{L}_k(\hat{\theta}_{k+1}),\, \hat{\theta}_{k+1} - \bar{\theta}\big\rangle \le \langle \hat{g}, \Delta_\theta\rangle \le \|\hat{g}\|\,\|\Delta_\theta\|,$$

where we have applied the first and second inequalities of equation 55 with $q = \hat{L}_k$, $\mu_q = \mu_\epsilon/2$ and $x^\star = \hat{\theta}_{k+1}$ in the first and second inequalities. Therefore, on $\mathcal{E}$,

$$\|\Delta_\theta\| \;\le\; \frac{2}{\mu_\epsilon}\,\|\hat{g}\|, \tag{70}$$

and hence $\|\hat{\theta}_{k+1} - \theta_k\| \le \|\Delta_\theta\| + \|\bar{\theta} - \theta_k\| \le 2t_b/\mu_\epsilon + c_r\eta/3 = 7c_r\eta/12$ by $\|\hat{g}\| \le t_b$, equation 65 and $t_b = \mu_\epsilon c_r\eta/8$. Thus $\hat{\theta}_{k+1}$ is interior, $\nabla\hat{L}_k(\hat{\theta}_{k+1}) = 0$, and on $\mathcal{E}$

$$\hat{H}_{\text{avg}}\,\Delta_\theta = \nabla\hat{L}_k(\bar{\theta}) - \nabla\hat{L}_k(\hat{\theta}_{k+1}) = \hat{g}, \tag{71}$$

$$\hat{H}_{\text{avg}} \succeq \frac{\mu_\epsilon}{2}\, I, \tag{72}$$

by the fundamental theorem of calculus and equation 60 on the segment from $\bar{\theta}$ to $\hat{\theta}_{k+1}$.

Expanding the batch maximizer. On $\mathcal{E}$ we know that

$$\|\Delta_{\text{avg}}\| \le \int_0^1 \left\|\hat{H}(\bar{\theta} + t\Delta_\theta) - \hat{H}(\bar{\theta})\right\| dt \le \frac{M_N}{2} \|\Delta_\theta\| \le \frac{M_N}{\mu_\epsilon} \|\hat{g}\|, \tag{73}$$

where we have applied equation 64 in the second inequality and equation 70 in the third inequality. Therefore, on $\mathcal{E}$ we have

$$\begin{aligned}\Delta_\theta = \hat{H}_{\text{avg}}^{-1}\hat{g} &= H^{-1}\hat{g} - H^{-1}(\Delta_H + \Delta_{\text{avg}})\,\hat{H}_{\text{avg}}^{-1}\hat{g} \\ &= \underbrace{H^{-1}\hat{g}}_{\mathcal{T}_1} - \underbrace{H^{-1}\Delta_H H^{-1}\hat{g}}_{\mathcal{T}_2} - \underbrace{H^{-1}\Delta_H\big(\hat{H}_{\text{avg}}^{-1} - H^{-1}\big)\hat{g}}_{\mathcal{T}_3} - \underbrace{H^{-1}\Delta_{\text{avg}}\,\Delta_\theta}_{\mathcal{T}_4},\end{aligned} \tag{74}$$

where we have applied equation 71 and equation 72 in the first equality, the resolvent identity with $\hat{H}_{\text{avg}} - H = \Delta_H + \Delta_{\text{avg}}$ in the second equality, and $\hat{H}_{\text{avg}}^{-1} = H^{-1} + (\hat{H}_{\text{avg}}^{-1} - H^{-1})$ with $\hat{H}_{\text{avg}}^{-1}\hat{g} = \Delta_\theta$ in the third equality. Since $\mathbb{E}\Delta_\theta = \mathbb{E}[\Delta_\theta\,\mathbb{I}_\mathcal{E}] + \mathbb{E}[\Delta_\theta\,\mathbb{I}_{\mathcal{E}^c}]$, it remains to bound $\|\mathbb{E}[\mathcal{T}_1\mathbb{I}_\mathcal{E}]\|$, $\|\mathbb{E}[\mathcal{T}_2\mathbb{I}_\mathcal{E}]\|$, $\mathbb{E}[(\|\mathcal{T}_3\| + \|\mathcal{T}_4\|)\mathbb{I}_\mathcal{E}]$ and $\|\mathbb{E}[\Delta_\theta\,\mathbb{I}_{\mathcal{E}^c}]\|$.

Bounding the leading term. From here on $N \ge N_0$, so that $\Pr(\mathcal{E}^c) \le 1/N$ and $N \ge \Lambda_d^2$. Since $\mathbb{E}\hat{g} = 0$ by equation 66, we know that

$$\left\|\mathbb{E}[\mathcal{T}_1\mathbb{I}_\mathcal{E}]\right\| = \left\|H^{-1}\,\mathbb{E}[\hat{g}\,\mathbb{I}_{\mathcal{E}^c}]\right\| \le \frac{1}{\mu_\epsilon}\big(\mathbb{E}\|\hat{g}\|^2\big)^{1/2}\Pr(\mathcal{E}^c)^{1/2} \le \frac{2\sqrt{G_{\ell,2}}}{a_0\mu_F}\cdot\frac{\epsilon}{N}, \tag{75}$$

where we have applied the Cauchy–Schwarz inequality in the first inequality, and equation 67 with $\mu_\epsilon = a_0\mu_F/(2\epsilon)$ in the second inequality.

Bounding the covariance term. First, by independence and equation 66, only the diagonal terms survive in

$$\mathbb{E}\big[\Delta_H H^{-1}\hat{g}\big] = \frac{1}{N^2}\sum_{i,j}\mathbb{E}\big[(H_i - H)H^{-1}g_j\big] = \frac{1}{N}\,\mathbb{E}\big[H_1 H^{-1}g_1\big].$$

Therefore,

$$\begin{aligned}\left\|H^{-1}\,\mathbb{E}\big[\Delta_H H^{-1}\hat{g}\big]\right\| &\le \frac{1}{N\mu_\epsilon^2}\big(\mathbb{E}\|H_1\|^2\big)^{1/2}\big(\mathbb{E}\|g_1\|^2\big)^{1/2} \le \frac{8\kappa_1\sqrt{G_{\ell,2}}}{(a_0\mu_F)^2}\cdot\frac{\epsilon}{N}, \\ \left\|H^{-1}\,\mathbb{E}\big[\Delta_H H^{-1}\hat{g}\,\mathbb{I}_{\mathcal{E}^c}\big]\right\| &\le \frac{1}{\mu_\epsilon^2}\big(\mathbb{E}\|\Delta_H\|^4\big)^{1/4}\big(\mathbb{E}\|\hat{g}\|^4\big)^{1/4}\Pr(\mathcal{E}^c)^{1/2} \le \frac{40\,\kappa_1 G_\ell}{(a_0\mu_F)^2}\cdot\frac{\epsilon}{N},\end{aligned}$$

where we have applied the Cauchy–Schwarz inequality and Hölder's inequality with exponents $(4, 4, 2)$ in the first inequalities, and equation 67 with $16\sqrt{6} \le 40$ in the second inequalities. Combining the two bounds with $\mathbb{E}[\mathcal{T}_2\mathbb{I}_\mathcal{E}] = H^{-1}\mathbb{E}[\Delta_H H^{-1}\hat{g}] - H^{-1}\mathbb{E}[\Delta_H H^{-1}\hat{g}\,\mathbb{I}_{\mathcal{E}^c}]$ yields

$$\left\|\mathbb{E}[\mathcal{T}_2\mathbb{I}_\mathcal{E}]\right\| \;\le\; \frac{\kappa_1\big(8\sqrt{G_{\ell,2}} + 40\,G_\ell\big)}{(a_0\mu_F)^2}\cdot\frac{\epsilon}{N}. \tag{76}$$

Bounding the higher-order terms. First, by equation 69, we know that

$$\begin{aligned}\mathbb{E}\|\Delta_H\|^4 &= \int_0^\infty 4t^3\,\Pr\big(\|\Delta_H\| > t\big)\,dt \\ &\le \int_0^\infty 4t^3 \min\left\{1, 2d\,e^{-Nt^2/(8b_H^2)}\right\}dt + \int_0^\infty 4t^3\min\left\{1, 2d\,e^{-Nt/(4b_H)}\right\}dt \\ &= \frac{64\,b_H^4\,(\Lambda_d^2+1)}{N^2} + \frac{256\,b_H^4\,(\Lambda_d^4 + 6\Lambda_d^2 + 8\Lambda_d + 9)}{N^4} \le \frac{6272\,b_H^4\Lambda_d^2}{N^2},\end{aligned} \tag{77}$$

where we have applied the layer-cake formula in the first equality, $\frac{t^2}{2(2b^2+bt)} \ge \min\big(\frac{t^2}{8b^2}, \frac{t}{4b}\big)$ and $\min\{1, x+y\} \le \min\{1,x\} + \min\{1,y\}$ in the first inequality, direct integration with $\log(2d) =$

$\Lambda_d - 1$ in the second equality, and $1 \le \Lambda_d^2 \le N$ in the second inequality. On the other hand, on $\mathcal{E}$ we have

$$\begin{aligned}\|\mathcal{T}_3\| + \|\mathcal{T}_4\| &\le \frac{2}{\mu_\epsilon^3}\|\Delta_H\|\big(\|\Delta_H\| + \|\Delta_{\text{avg}}\|\big)\|\hat{g}\| + \frac{M_N}{2\mu_\epsilon}\|\Delta_\theta\|^2 \\ &\le \frac{2}{\mu_\epsilon^3}\|\Delta_H\|^2\,\|\hat{g}\| + \frac{5}{2\mu_\epsilon^3}\Big(\frac{C_3}{\epsilon} + 2M_w A_\psi\Big)\|\hat{g}\|^2,\end{aligned}$$

where we have applied $\hat{H}_{\text{avg}}^{-1} - H^{-1} = -H^{-1}(\Delta_H + \Delta_{\text{avg}})\hat{H}_{\text{avg}}^{-1}$, $\|H^{-1}\| \le 1/\mu_\epsilon$, $\|\hat{H}_{\text{avg}}^{-1}\| \le 2/\mu_\epsilon$ and equation 73 in the first inequality, and $\|\Delta_H\| \le \mu_\epsilon/4$, equation 73, equation 70 and $m_N \le 2A_\psi$ in the second inequality. Taking expectation on $\mathcal{E}$ yields

$$\begin{aligned}\mathbb{E}\big[(\|\mathcal{T}_3\| + \|\mathcal{T}_4\|)\,\mathbb{I}_\mathcal{E}\big] &\le \frac{2}{\mu_\epsilon^3}\big(\mathbb{E}\|\Delta_H\|^4\big)^{1/2}\big(\mathbb{E}\|\hat{g}\|^2\big)^{1/2} + \frac{5}{2\mu_\epsilon^3}\Big(\frac{C_3}{\epsilon} + 2M_w A_\psi\Big)\mathbb{E}\|\hat{g}\|^2 \\ &\le \frac{2}{\mu_\epsilon^3}\cdot\frac{80\,b_H^2\Lambda_d}{N}\cdot\frac{\sqrt{G_{\ell,2}}}{\sqrt{N}} + \frac{5}{2\mu_\epsilon^3}\Big(\frac{C_3}{\epsilon} + 2M_w A_\psi\Big)\frac{G_{\ell,2}}{N} \\ &= \frac{20480\,\kappa_1^2\Lambda_d\sqrt{G_{\ell,2}}}{(a_0\mu_F)^3\sqrt{N}}\cdot\frac{\epsilon}{N} + \frac{20\,\epsilon\,\big(C_3 + 2M_w A_\psi\epsilon\big)G_{\ell,2}}{(a_0\mu_F)^3}\cdot\frac{\epsilon}{N} \\ &\le \frac{20480\,\kappa_1^2\sqrt{G_{\ell,2}} + 20\,\big(C_3 + 2M_w A_\psi\big)G_{\ell,2}}{(a_0\mu_F)^3}\cdot\frac{\epsilon}{N},\end{aligned} \tag{78}$$

where we have applied the Cauchy–Schwarz inequality in the first inequality, equation 77 with $\sqrt{6272} \le 80$ and equation 67 in the second inequality, $b_H = 4\kappa_1/\epsilon$ in the equality, and $N \ge \Lambda_d^2$ and $\epsilon \le 1$ in the third inequality.

Bounding the bad event. Since $\hat{\theta}_{k+1}, \bar{\theta} \in \mathcal{B}_k$ surely, we know that

$$\big\|\mathbb{E}[\Delta_\theta\,\mathbb{I}_{\mathcal{E}^c}]\big\| \le 2c_r\eta\,\Pr(\mathcal{E}^c) \le \frac{2c_r(1-\gamma)}{a_0}\cdot\frac{\epsilon}{N}. \tag{79}$$

Bounding the bias. Combining equation 74 with equation 75, equation 76, equation 78 and equation 79 yields

$$\big\|\mathbb{E}[\hat{\theta}_{k+1}] - \bar{\theta}\big\| \le \big\|\mathbb{E}[\mathcal{T}_1\mathbb{I}_\mathcal{E}]\big\| + \big\|\mathbb{E}[\mathcal{T}_2\mathbb{I}_\mathcal{E}]\big\| + \mathbb{E}\big[(\|\mathcal{T}_3\| + \|\mathcal{T}_4\|)\,\mathbb{I}_\mathcal{E}\big] + \big\|\mathbb{E}[\Delta_\theta\,\mathbb{I}_{\mathcal{E}^c}]\big\| \le \frac{C_b\,\epsilon}{N},$$

with

$$\begin{aligned}C_b = \frac{2\sqrt{G_{\ell,2}}}{a_0\mu_F} + \frac{\kappa_1\big(8\sqrt{G_{\ell,2}} + 40\,G_\ell\big)}{(a_0\mu_F)^2} \\ + \frac{20480\,\kappa_1^2\sqrt{G_{\ell,2}} + 20\,\big(C_3 + 2M_w A_\psi\big)G_{\ell,2}}{(a_0\mu_F)^3} + \frac{2c_r(1-\gamma)}{a_0}.\end{aligned} \tag{80}$$

Dividing by $\eta = \epsilon(1-\gamma)/a_0$ gives the bound C_N/N in direction units.

Bounding the variance. Finally, since $\mathrm{Var}(X \mid \theta_k) \le \mathbb{E}\|X - c\|^2$ for every deterministic c, taking $c = (\bar{\theta} - \theta_k)/\eta$ gives

$$\mathrm{Var}(\hat{\omega}_k \mid \theta_k) \le \frac{1}{\eta^2}\Big(\mathbb{E}\big[\|\Delta_\theta\|^2\,\mathbb{I}_\mathcal{E}\big] + 4c_r^2\eta^2\,\Pr(\mathcal{E}^c)\Big) \le \frac{1}{\eta^2}\Big(\frac{4}{\mu_\epsilon^2}\cdot\frac{G_{\ell,2}}{N} + \frac{4c_r^2\eta^2}{N}\Big) = \frac{\sigma_\omega^2}{N},$$

where we have applied $\|\Delta_\theta\| \le 2c_r\eta$ in the first inequality, equation 70 and equation 67 in the second inequality, and $\mu_\epsilon\eta = \mu_F(1-\gamma)/2$ in the equality, with

$$\sigma_\omega^2 = \frac{16\,G_{\ell,2}}{\mu_F^2(1-\gamma)^2} + 4c_r^2. \tag{81}$$

□

The batch enters at first order only through $\mathcal{T}_1$, which has mean zero because $\hat{L}_k$ is unbiased for $\bar{L}_k$, so restricting to $\mathcal{E}$ adds only a factor $\sqrt{p_N}$. The term $\mathcal{T}_2$ contributes only through its diagonal covariance, which carries $1/N$, and $\mathcal{T}_3$ and $\mathcal{T}_4$ are quadratic in the fluctuations. Hence C_b is free of ϵ, N and d, which enters only through N_0. The two-sided bound equation 60 also gives the inner loop of Lemma D.4 a condition number free of ϵ and N.

D.5 INNER LOOP AND ASSEMBLY

Lemma D.4 (Inner loop). *Let $\epsilon > 0$, let $\mathcal{B}_k = \{\theta' : \|\theta' - \theta_k\| \le c_r\eta\}$ be the trust ball and $\Pi_{\mathcal{B}_k}$ the Euclidean projection onto it, let $\hat{L}_k$ be the batch objective of equation 15, let $\hat{\theta}_{k+1}$ be a maximizer of $\hat{L}_k$ on $\mathcal{B}_k$, and, with $\kappa_1 = 8a_0G^2$ of equation 56, set*

$$\mu_{\text{in}} = \frac{a_0\mu_F}{4\epsilon}, \qquad L_{\text{in}} = \frac{5\kappa_1}{3\epsilon}, \qquad h_{\text{in}} = \frac{2}{\mu_{\text{in}} + L_{\text{in}}}, \qquad c_0 = \frac{\mu_{\text{in}}}{L_{\text{in}}} = \frac{3\mu_F}{160\,G^2}. \tag{82}$$

Suppose that $\mu_{\text{in}} I \preceq -\nabla^2\hat{L}_k(\theta') \preceq L_{\text{in}} I$ for every $\theta' \in \mathcal{B}_k$. Then the iterates $\theta^{(t+1)} = \Pi_{\mathcal{B}_k}\big(\theta^{(t)} + h_{\text{in}}\nabla\hat{L}_k(\theta^{(t)})\big)$, $t = 0, \dots, T-1$, of projected gradient ascent from $\theta^{(0)} = \theta_k$ satisfy

$$\frac{\big\|\theta^{(T)} - \hat{\theta}_{k+1}\big\|}{\eta} \;\le\; (1 - c_0)^T c_r \;\le\; c_r e^{-c_0 T}. \tag{83}$$

Proof. The projection onto $\mathcal{B}_k$ is nonexpansive (Nesterov, 2018, Corollary 2.2.3), and $\hat{\theta}_{k+1} = \Pi_{\mathcal{B}_k}\big(\hat{\theta}_{k+1} + h_{\text{in}}\nabla\hat{L}_k(\hat{\theta}_{k+1})\big)$ since $\hat{L}_k$ is twice continuously differentiable on $\mathbb{R}^d$ and $-\nabla^2\hat{L}_k \succeq \mu_{\text{in}} I$ on $\mathcal{B}_k$, which has nonempty interior because $c_r\eta > 0$ (Nesterov, 2018, Theorem 2.2.12). Every iterate lies in $\mathcal{B}_k$, and hence so does the segment from $\hat{\theta}_{k+1}$ to $\theta^{(t)}$. Therefore, for $t = 0, \dots, T-1$ we have

$$\begin{aligned}
\big\|\theta^{(t+1)} - \hat{\theta}_{k+1}\big\| &\le \Big\|\theta^{(t)} - \hat{\theta}_{k+1} + h_{\text{in}}\big(\nabla\hat{L}_k(\theta^{(t)}) - \nabla\hat{L}_k(\hat{\theta}_{k+1})\big)\Big\| \\
&= \big\|(I - h_{\text{in}}\hat{H}^{(t)}_{\text{avg}})(\theta^{(t)} - \hat{\theta}_{k+1})\big\| \le \frac{L_{\text{in}} - \mu_{\text{in}}}{L_{\text{in}} + \mu_{\text{in}}}\big\|\theta^{(t)} - \hat{\theta}_{k+1}\big\| \\
&\le (1 - c_0)\big\|\theta^{(t)} - \hat{\theta}_{k+1}\big\|,
\end{aligned}$$

where $\hat{H}^{(t)}_{\text{avg}} = -\int_0^1 \nabla^2\hat{L}_k\big(\hat{\theta}_{k+1} + s(\theta^{(t)} - \hat{\theta}_{k+1})\big)\,ds$ is symmetric with $\mu_{\text{in}} I \preceq \hat{H}^{(t)}_{\text{avg}} \preceq L_{\text{in}} I$, and we have applied nonexpansiveness and the fixed-point identity in the first inequality, the fundamental theorem of calculus in the equality, $\|I - h_{\text{in}}\hat{H}^{(t)}_{\text{avg}}\| \le \max_{\lambda\in[\mu_{\text{in}}, L_{\text{in}}]}|1 - h_{\text{in}}\lambda|$ with $h_{\text{in}} = 2/(\mu_{\text{in}} + L_{\text{in}})$ in the second inequality (Nesterov, 2018, cf. Theorem 2.2.14 and (2.2.62)), and $\mu_{\text{in}} \le L_{\text{in}}$, which holds since $\mu_F \le G^2$, in the third inequality. Iterating from $\theta^{(0)} = \theta_k$ and using $\|\theta_k - \hat{\theta}_{k+1}\| \le c_r\eta$ and $0 \le 1 - c_0 \le e^{-c_0}$ gives equation 83. □

The hypothesis of the lemma is what the event $\mathcal{E}$ of Lemma D.3 provides through equation 60. The condition number $L_{\text{in}}/\mu_{\text{in}} = 160G^2/(3\mu_F)$ is free of ϵ, N and the advantage scale, because for $\epsilon \le \epsilon_0$ the advantage part $2\kappa_2 A_\psi$ of the batch curvature is dominated by κ_1/ϵ.

Proposition D.5 (Global convergence). *Suppose Assumptions 3–5, $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ fixed, $\epsilon \le \epsilon_0$ and $N \ge N_0$, with ϵ_0 and N_0 of equation 57 and equation 56, and let $\theta_{k+1} = \theta^{(T)}$ be computed with the step size h_{in} of Lemma D.4. Then with $\omega_k = (\theta_{k+1} - \theta_k)/\eta$ and $\omega_k^* = F_\rho(\theta_k)^\dagger\nabla J(\theta_k)$ we have*

$$J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[J(\theta_k)\big] \;\le\; \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} \;+\; G\Big[C_{\text{prox}}(\alpha)\,\eta + \frac{C_N}{N} + c_r e^{-c_0 T}\Big] \;+\; \frac{Bc_r^2}{2}\,\eta \;+\; \frac{\mathrm{KL}_0}{\eta K}, \tag{84}$$

where $C_{\text{prox}}(\alpha) \le C\,(1 + |\alpha - 2|)$, and C_N and the inverse condition number c_0 of Lemma D.4 are free of ϵ.

For a fixed ϵ, equation 84 leaves a gap of $\sqrt{\epsilon_{\text{bias}}}/(1-\gamma) + \mathcal{O}(\epsilon)$ as $K \to \infty$.

Proof of Proposition D.5. Let $\epsilon \le \epsilon_0$ with ϵ_0 of equation 57, $N \ge N_0$ with N_0 of equation 56, and $\theta_{k+1} = \theta^{(T)}$ as in Lemma D.4. We apply Lemma D.1 to $\omega_k = (\theta_{k+1} - \theta_k)/\eta$, which requires bounds on $\mathbb{E}\|\omega_k\|^2$ and $\mathbb{E}\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\|$.

Bounding the second moment. Since every inner iterate lies in $\mathcal{B}_k$, surely

$$\|\omega_k\| = \frac{\big\|\theta^{(T)} - \theta_k\big\|}{\eta} \le c_r. \tag{85}$$

Bounding the conditional bias. Let $\mathcal{E}$ be the event equation 58 at iteration k, $\bar{\theta}_{k+1}$ the population maximizer of Lemma D.2, and $\hat{\theta}_{k+1}$ the maximizer of $\hat{L}_k$ on $\mathcal{B}_k$ on $\mathcal{E}$ and $\theta^{(T)}$ off $\mathcal{E}$. This choice is admissible in Lemma D.3, since $\theta^{(T)} \in \mathcal{B}_k$ is a Borel function of θ_k and the batch, and so is the maximizer on $\mathcal{E}$ by the proof of that lemma. Since the batch of iteration k is drawn afresh given θ_k, Lemma D.3 applies conditionally on θ_k, and on $\mathcal{E}$, equation 60 gives the hypothesis of Lemma D.4, with $\mu_\epsilon/2 = \mu_{\text{in}}$ and $\kappa_1/\epsilon + 2\kappa_2 A_\psi \le L_{\text{in}}$ by the sixth entry of equation 57. Therefore,

$$\begin{aligned}
\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\| &= \Big\|\frac{\bar{\theta}_{k+1} - \theta_k}{\eta} - \omega_k^* + \frac{\mathbb{E}[\hat{\theta}_{k+1} \mid \theta_k] - \bar{\theta}_{k+1}}{\eta} + \frac{\mathbb{E}\big[(\theta^{(T)} - \hat{\theta}_{k+1})\,\mathbb{I}_{\mathcal{E}} \mid \theta_k\big]}{\eta}\Big\| \\
&\le \Big\|\frac{\bar{\theta}_{k+1} - \theta_k}{\eta} - \omega_k^*\Big\| + \frac{\big\|\mathbb{E}[\hat{\theta}_{k+1} \mid \theta_k] - \bar{\theta}_{k+1}\big\|}{\eta} + \frac{\mathbb{E}\big[\|\theta^{(T)} - \hat{\theta}_{k+1}\|\,\mathbb{I}_{\mathcal{E}} \mid \theta_k\big]}{\eta} \\
&\le C_{\text{prox}}(\alpha)\,\eta + \frac{C_N}{N} + c_r e^{-c_0 T},
\end{aligned} \tag{86}$$

where we have applied $\theta^{(T)} - \hat{\theta}_{k+1} = (\theta^{(T)} - \hat{\theta}_{k+1})\,\mathbb{I}_{\mathcal{E}}$ and the $\sigma(\theta_k)$-measurability of $\bar{\theta}_{k+1}$ in the equality, the triangle and Jensen inequalities in the first inequality, and equation 31, equation 61 with $\epsilon/\eta = a_0/(1-\gamma)$, and equation 83 on $\mathcal{E}$ in the second inequality.

Finally, combining equation 85 and equation 86 with Lemma D.1 yields

$$\begin{aligned}
J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[J(\theta_k)\big] &\le \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + \frac{G}{K}\sum_{k=0}^{K-1}\mathbb{E}\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\| + \frac{B\eta}{2K}\sum_{k=0}^{K-1}\mathbb{E}\|\omega_k\|^2 + \frac{\text{KL}_0}{\eta K} \\
&\le \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} + G\Big[C_{\text{prox}}(\alpha)\,\eta + \frac{C_N}{N} + c_r e^{-c_0 T}\Big] + \frac{B c_r^2}{2}\,\eta + \frac{\text{KL}_0}{\eta K},
\end{aligned}$$

which is equation 84, with $C_{\text{prox}}(\alpha) \le \max\big(c_1, G^3 G_\omega^2/(2\mu_F)\big)(1 + |\alpha - 2|)$ by equation 31, C_N free of ϵ, and $c_0 = 3\mu_F/(160 G^2)$. □

Setting $\hat{\theta}_{k+1} = \theta^{(T)}$ off $\mathcal{E}$ counts the bad event once, in equation 61, so the inner loop adds no statistical term beyond C_N/N.

Proof of Theorem 4.2. We fix the parameters, check the hypotheses of Proposition D.5, bound each term of equation 84 by $\bar{\varepsilon}/4$, and bound the parameters.

Fixing the parameters. Take $C_{\text{prox}}(\alpha) = c_1 + \frac{G^3 G_\omega^2}{2\mu_F}|\alpha - 2|$, which is admissible in equation 31, and write $C_\eta = G\,C_{\text{prox}}(\alpha) + B c_r^2/2$. We choose

$$\begin{aligned}
\eta &= \frac{\bar{\varepsilon}}{4\big(G\,C_{\text{prox}}(\alpha) + B c_r^2/2\big)}, & K &= \Big\lceil \frac{4\,\text{KL}_0}{\eta\,\bar{\varepsilon}} \Big\rceil, \\
N &= \Big\lceil \frac{4\,G\,C_N}{\bar{\varepsilon}} \Big\rceil, & T &= \Big\lceil \frac{1}{c_0}\log\frac{4\,G\,c_r}{\bar{\varepsilon}} \Big\rceil,
\end{aligned} \tag{87}$$

with $\epsilon = a_0\eta/(1-\gamma)$, h_{in} and c_0 of equation 82, and $0 < \bar{\varepsilon} \le \bar{\varepsilon}_0$, where

$$\bar{\varepsilon}_0 = \min\Big\{\frac{4(1-\gamma)\,C_\eta\,\epsilon_0}{a_0},\ \frac{4\,G\,C_N}{N_0},\ \frac{R}{e},\ 4\sqrt{\text{KL}_0\,C_\eta}\Big\}, \tag{88}$$

which is positive if and only if $\text{KL}_0 > 0$.

Verifying the hypotheses. First, by equation 87 and equation 88, we have

$$\epsilon = \frac{a_0\,\bar{\varepsilon}}{4(1-\gamma)\,C_\eta} \le \epsilon_0, \qquad N \ge \frac{4\,G\,C_N}{\bar{\varepsilon}} \ge N_0, \qquad K \ge \frac{4\,\text{KL}_0}{\eta\,\bar{\varepsilon}} = \frac{16\,\text{KL}_0\,C_\eta}{\bar{\varepsilon}^2} \ge 1,$$

where we have applied $\eta = \bar{\varepsilon}/(4C_\eta)$ in the equalities, and the first, second and fourth entries of equation 88 in the corresponding last inequalities. Hence Proposition D.5 applies.

Bounding the four terms. Next, by equation 84, we have

$$\begin{aligned}
J^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big[J(\theta_k)\big] - \frac{\sqrt{\epsilon_{\text{bias}}}}{1-\gamma} &\le C_\eta\,\eta + \frac{G\,C_N}{N} + G\,c_r e^{-c_0 T} + \frac{\text{KL}_0}{\eta K} \le \frac{\bar{\varepsilon}}{4} + \frac{\bar{\varepsilon}}{4} + \frac{\bar{\varepsilon}}{4} + \frac{\bar{\varepsilon}}{4} \\
&= \bar{\varepsilon},
\end{aligned}$$

where the second inequality holds term by term by equation 87. This is the accuracy claimed in Theorem 4.2.

Bounding the parameters. This step collects the explicit constants and the discount tally. Throughout, $a_0 = c_a R/(1-\gamma)$ with $c_a > 0$ free of R and γ, and $c_A = 4$.

Advantage scale. Each sample of Agarwal et al. (2021, Algorithm 3) runs two consecutive rollouts of π_{θ_k} with independent geometric lengths. The first returns $(s_i, a_i) \sim \nu_\rho^{\pi_{\theta_k}}$. The second collects T_i+1 rewards with $\Pr(T_i = t) = (1-\gamma)\gamma^t$, starts at s_i with the action a_i or a fresh action with probability $1/2$ each, and returns $\hat{A}_i = \pm 2 \sum_{j=0}^{T_i} r(s_j, a_j)$, which satisfies $\mathbb{E}[\hat{A}_i \mid s_i, a_i] = A^{\pi_{\theta_k}}(s_i, a_i)$. Since rewards lie in $[-R, R]$, we have

$$\begin{aligned} \|\hat{A}_i\|_{\psi_1} &\le 2R\,\|T_i + 1\|_{\psi_1} = \frac{2R}{\log\big(2/(1+\gamma)\big)} \le \frac{c_A R}{1-\gamma} = A_\psi, \\ \mathbb{E}\big[\hat{A}_i^2\big] &\le 4R^2\,\mathbb{E}(T_i+1)^2 = \frac{4(1+\gamma)R^2}{(1-\gamma)^2} \le \frac{c_A^2 R^2}{(1-\gamma)^2} = A_2, \end{aligned} \tag{89}$$

where we have applied $|\hat{A}_i| \le 2R(T_i+1)$ in the first inequality of each row, $\mathbb{E}\exp((T_i+1)/c) = (1-\gamma)e^{1/c}/(1-\gamma e^{1/c})$, which equals 2 at $c = 1/\log(2/(1+\gamma))$, in the first equality of the first row, $\log(2/(1+\gamma)) \ge (1-\gamma)/2$ in its second inequality, $\mathbb{E}(T_i+1)^2 = (1+\gamma)/(1-\gamma)^2$ in the first equality of the second row, and $4(1+\gamma) \le c_A^2$ in its second inequality; the last equalities fix A_ψ and A_2. Each sample costs at most $2/(1-\gamma)$ environment steps in expectation. In the same way, $\mathbb{E}|\hat{A}_i| \le 2R\,\mathbb{E}(T_i+1) = 2R/(1-\gamma)$, so $a_0 = c_a R/(1-\gamma)$ is of the order of this bound on the batch scale $\bar{a} = \mathbb{E}|\hat{A}|$ of the uniform-weight penalty of Lemma C.4, which is why we adopt it.

Thresholds. With equation 89, $c_w = 2Gc_r(1-\gamma)/a_0$ of Appendix D.1, $c_r = 4GR/((1-\gamma)^2\mu_F)$ and $A_\infty \le 2R/(1-\gamma)$, we have

$$\begin{aligned} &c_w = \frac{8G^2}{c_a \mu_F}, \qquad \frac{C_{\mathrm{dev}}}{a_0} \le \frac{1}{c_a}\Big(\frac{G^2\big((18+9c_{\max})G^2 + 14B\big)}{\mu_F} + 2(G^2+B)\Big), \qquad \frac{\kappa_1}{A_\psi} = 2c_a G^2, \\ &\frac{c_r(1-\gamma)\,(2\bar{C}_3 + 3M_w A_\psi)}{a_0^2} = \frac{4G\,\big(36c_a(1+c_{\max}) + 48\big)(G^3 + GB + B_2)}{c_a^2\,\mu_F}, \end{aligned} \tag{90}$$

so every entry of equation 32 and equation 57 is bounded below by a positive function of c_a, $c_{\max}$, G, B, B_2 and μ_F alone, and so is ϵ_0. Moreover, $t_b = GR/(4(1-\gamma))$ and $G_\ell = 2G(c_A + 24G^2/\mu_F)R/(1-\gamma)$ give $t_b/G_\ell = 1/(32 + 192G^2/\mu_F) \le 1$, so that

$$\begin{aligned} c_b &= \frac{1}{6\,(32 + 192G^2/\mu_F)^2} \ge \frac{\mu_F^2}{301056\,G^4} \ge c_H, \qquad c_\star = c_H = \frac{\mu_F^2}{327680\,G^4}, \\ N_0 &= \max\Big(\Big\lceil \frac{2}{c_H} \log \frac{8d}{c_H} \Big\rceil,\ \lceil \Lambda_d^2 \rceil\Big), \end{aligned} \tag{91}$$

where we have applied $G^2/\mu_F \ge 1$ in the first inequality, $301056 \le 327680$ in the second inequality, and $c_H \le 1/20$ in $c_\star = c_H$. Hence N_0 depends only on G, μ_F and d.

Step, batch and accuracy constants. First, we have

$$C_\eta \ge G\,c_1 + \frac{G^4 G_\omega^2}{2\mu_F}\,|\alpha - 2| \ge \frac{G^4 G_\omega^2}{2\mu_F}\,\big(1 + |\alpha - 2|\big) = \frac{2\,G^6 R^2\,(1 + |\alpha - 2|)}{(1-\gamma)^4\,\mu_F^3}, \tag{92}$$

where we have applied $B c_r^2 \ge 0$ and the choice of $C_{\mathrm{prox}}(\alpha)$ in the first inequality, $c_1 \ge (G^2 + B)\,G\,G_\omega^2/\mu_F \ge G^3 G_\omega^2/(2\mu_F)$ by equation 33 in the second inequality, and $G_\omega = 2GR/((1-\gamma)^2\mu_F)$ in the equality. On the other hand, with $c_\eta = (10G^6 + 8G^4 B)/\mu_F^3 + 8BG^2/\mu_F^2$, $g_1 = 2\sqrt{2}\,G\,(c_A + 24G^2/\mu_F)$ and

$$c_{\mathrm{st}} = \frac{1}{c_a \mu_F}\Big(2g_1 + 8G + \frac{384\,G^2 g_1}{\mu_F} + \frac{1310720\,G^4 g_1}{\mu_F^2} + \frac{20\,(18c_a + 8c_A)(G^3 + GB + B_2)\,g_1^2}{c_a^2 \mu_F^2}\Big),$$

we know that

$$C_\eta \le \Big(\frac{2G^2(G^2+B)\,G_\omega^2}{\mu_F} + \frac{G^4 G_\omega^2}{2\mu_F} + 2BG_\omega^2\Big)\big(1+|\alpha-2|\big) = c_\eta\,\frac{(1+|\alpha-2|)\,R^2}{(1-\gamma)^4}, \tag{93}$$

$$C_N = \frac{c_a\,C_b\,R}{(1-\gamma)^2} \le c_a\,c_{\mathrm{st}}\,\frac{(1+|\alpha-2|)\,R}{(1-\gamma)^2}, \tag{94}$$

where we have applied the choice of $C_{\mathrm{prox}}(\alpha)$, $c_1 \le 2G(G^2+B)\,G_\omega^2/\mu_F$ by equation 33, $c_r = 2G_\omega$ and $1 \le 1+|\alpha-2|$ in the first inequality, the definition of G_ω in the first equality, and $C_N = C_b\,a_0/(1-\gamma)$ in the second equality. The last inequality follows from equation 80 term by term, using $a_0 c_w = 8G^2R/((1-\gamma)\mu_F)$, $G_\ell, \sqrt{G_{\ell,2}} \le g_1 R/(1-\gamma)$ by equation 89, $\kappa_1 = 8c_a G^2 R/(1-\gamma)$, $C_3 + 2M_w A_\psi \le (18c_a + 8c_A)(1+|\alpha-2|)(G^3+GB+B_2)R/(1-\gamma)$ and $2c_r(1-\gamma)/a_0 = 8G/(c_a\mu_F)$. In the other direction, the bad-event term of equation 80 gives $C_N \ge 2c_r = 8GR/((1-\gamma)^2\mu_F)$. Combining this and equation 92 with equation 88 yields

$$\bar\varepsilon_0 \;\ge\; R\,\min\Big\{\frac{8\,G^6\epsilon_0}{c_a\,\mu_F^3},\ \frac{32\,G^2}{\mu_F N_0},\ \frac{1}{e},\ \frac{4\sqrt{2}\,G^3\sqrt{\mathrm{KL}_0}}{\mu_F^{3/2}}\Big\}, \tag{95}$$

where we have used $1-\gamma \le 1$ and $1+|\alpha-2| \ge 1$.

Discount tally. By equation 87, we have

$$\begin{aligned}
\eta &= \frac{\bar\varepsilon}{4\,C_\eta} \le \frac{\mu_F^3}{8\,G^6}\cdot\frac{(1-\gamma)^4\,\bar\varepsilon}{(1+|\alpha-2|)\,R^2},\\
\epsilon &= \frac{a_0\,\bar\varepsilon}{4(1-\gamma)\,C_\eta} \le \frac{c_a\,\mu_F^3}{8\,G^6}\cdot\frac{(1-\gamma)^2\,\bar\varepsilon}{(1+|\alpha-2|)\,R},\\
N &\le \frac{4\,G\,C_N}{\bar\varepsilon} + 1 \le \frac{8\,G\,C_N}{\bar\varepsilon} \le 8\,G\,c_a\,c_{\mathrm{st}}\,\frac{(1+|\alpha-2|)\,R}{(1-\gamma)^2\,\bar\varepsilon},\\
K &\le \frac{16\,\mathrm{KL}_0\,C_\eta}{\bar\varepsilon^2} + 1 \le \frac{32\,\mathrm{KL}_0\,C_\eta}{\bar\varepsilon^2} \le 32\,c_\eta\,\frac{\mathrm{KL}_0\,(1+|\alpha-2|)\,R^2}{(1-\gamma)^4\,\bar\varepsilon^2},\\
2NK &\le \frac{512\,G\,\mathrm{KL}_0\,C_N C_\eta}{\bar\varepsilon^3} \le 512\,G\,c_a\,c_{\mathrm{st}}\,c_\eta\,\frac{\mathrm{KL}_0\,(1+|\alpha-2|)^2\,R^3}{(1-\gamma)^6\,\bar\varepsilon^3},
\end{aligned}$$

where we have applied equation 92 for η, equation 92 with $a_0 = c_a R/(1-\gamma)$ for ϵ, $\lceil x\rceil \le x+1$ with $4GC_N/\bar\varepsilon \ge 1$ and $16\,\mathrm{KL}_0 C_\eta/\bar\varepsilon^2 \ge 1$ for N and K, the bounds on N and K in the first inequality for $2NK$, and equation 94 and equation 93 in the last inequalities. Finally, for the inner loop we know that

$$\begin{aligned}
T &\le 1 + \frac{160\,G^2}{3\mu_F}\log\frac{16\,G^2 R}{(1-\gamma)^2\mu_F\,\bar\varepsilon} \le 1 + \frac{160\,G^2}{3\mu_F}\Big(\log\frac{16\,G^2}{\mu_F} + 2\log\frac{R}{(1-\gamma)\,\bar\varepsilon}\Big)\\
&\le \Big(1 + \frac{160\,G^2}{3\mu_F}\Big(2 + \log\frac{16\,G^2}{\mu_F}\Big)\Big)\log\frac{R}{(1-\gamma)\,\bar\varepsilon},
\end{aligned}$$

where we have applied $\lceil x\rceil \le x+1$, $c_0 = 3\mu_F/(160G^2)$ and $4Gc_r = 16G^2R/((1-\gamma)^2\mu_F)$ in the first inequality, $\bar\varepsilon \le R$ in the second inequality, and $\log\frac{R}{(1-\gamma)\bar\varepsilon} \ge 1$ for $\bar\varepsilon \le R/e$ with $\log(16G^2/\mu_F) \ge 0$ in the third inequality. Each sample uses two trajectories and the T inner steps reuse the batch, so the algorithm samples $2NK$ trajectories and at most $2NK/(1-\gamma)$ environment steps in expectation, which gives the orders in Theorem 4.2. □

With $a_0 \propto R/(1-\gamma)$, the discount enters the tally only through $c_r \propto (1-\gamma)^{-2}$ and the advantage scales $A_\psi, A_\infty \propto (1-\gamma)^{-1}$ and $A_2 \propto (1-\gamma)^{-2}$, while c_w, c_0, N_0 and the lower bound on ϵ_0 are free of R and γ. Thus $C_\eta \propto (1-\gamma)^{-4}$ sets $\eta \propto (1-\gamma)^4$ and $K \propto (1-\gamma)^{-4}$, $C_N \propto (1-\gamma)^{-2}$ sets $N \propto (1-\gamma)^{-2}$, and each sample adds $(1-\gamma)^{-1}$ in environment steps, for $\mathcal{O}\big((1-\gamma)^{-7}\bar\varepsilon^{-3}\big)$ steps in total; by equation 95, the admissible accuracy range scales with R and, for fixed KL_0, does not shrink as $\gamma \to 1$. The term $|\alpha-2|$ enters twice, through $C_{\mathrm{prox}}(\alpha)$ in K and through C_3 in N, and over $[\alpha_{\min}, \alpha_{\max}]$ both factors are at most $1 + c_{\max}$. The trajectory count carries no logarithm, which enters only through T, the number of gradient steps on the stored batch.

Proof that Theorem 4.2 covers adaptive update rules. Let $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ be measurable with respect to the σ-field $\mathcal{F}_k$ generated by θ_0 and the batches of iterations $0, \dots, k-1$, so that $\sigma(\theta_k) \subseteq \mathcal{F}_k$ and, given $\mathcal{F}_k$, the batch of iteration k has the law of Lemma D.3 at θ_k. Since f_α, f'_α and f''_α are continuous in (α, w), $\bar{L}_k$, $\hat{L}_k$ and their first two derivatives are continuous in α by dominated convergence and Borel in the other arguments, hence jointly Borel, and the measurability construction in the proof of Lemma D.3, run with α as an extra argument, makes $\bar{\theta}_{k+1}$, $\mathcal{E}$, $\hat{\theta}_{k+1}$ and $\theta^{(T)}$ Borel in (θ_k, α_k) and the batch. The quantities ϵ_{prox}, ϵ_0, N_0, h_{in}, c_0 and the constants of equation 56 other than C_3 depend on α only through $c_{\max}$, and C_b is increasing in $C_3 \le \bar{C}_3$. Hence, conditioning equation 86 on $\mathcal{F}_k$,

$$\mathbb{E}\big\|\mathbb{E}[\omega_k \mid \theta_k] - \omega_k^*\big\| = \mathbb{E}\big\|\mathbb{E}\big[\mathbb{E}[\omega_k \mid \mathcal{F}_k] - \omega_k^* \,\big|\, \theta_k\big]\big\| \le \mathbb{E}\big\|\mathbb{E}[\omega_k \mid \mathcal{F}_k] - \omega_k^*\big\|$$
$$\le \bar{C}_{\text{prox}}\,\eta + \frac{\bar{C}_N}{N} + c_r e^{-c_0 T},$$

where $\bar{C}_{\text{prox}} = c_1 + G^3 G_\omega^2 c_{\max}/(2\mu_F)$ and $\bar{C}_N$ is C_N with $\bar{C}_3$ in place of C_3, and we have applied the tower property in the equality, Jensen's inequality in the first inequality, and equation 31, equation 61 and equation 83 conditionally on $\mathcal{F}_k$ in the second inequality. The proofs of Proposition D.5 and Theorem 4.2 then apply with $\bar{C}_{\text{prox}}$, $\bar{C}_N$ and $c_{\max}$ in place of $C_{\text{prox}}(\alpha)$, C_N and $|\alpha - 2|$, since $\bar{C}_N \le c_a c_{\text{st}}(1 + c_{\max})R/(1-\gamma)^2$ as in equation 94. □

E PROCGEN EXPERIMENT DETAILS

Both experiments with $\epsilon \in \{0.2, 0.5\}$ use the easy-mode suite by Cobbe et al. (2020). Table 3 lists every hyperparameter used in the experiments. Every environment, optimiser, and network setting is common between methods. Evaluation draws episodes from the full level distribution, which the training levels overlap with probability $200/2^{31}$. The backbone uses the framework default initialisation; orthogonal gain $\sqrt{2}$ on every convolution of a residual stack compounds across the six blocks.

At $\epsilon = 0.5$ the original α^{SNR} update rule collapsed, so we cap the ceiling at $\alpha_{\max} = 2$. The α^{SNR} column of Table 2 uses that tuned update rule. We fitted the two variants on BossFight and BigFish at seed 1.

F MUJOCO EXPERIMENT DETAILS

Table 5 lists every hyperparameter of the MuJoCo experiments used in Section 5.2.

The learning rate is tuned per environment: Pusher and Swimmer use 2×10^{-4} and the remaining six environments use 5×10^{-4}. We selected it from the two-point sweep on Reacher and Swimmer, and the learning rate is the same for every method in an environment to keep comparisons level.

F.1 TWO-LAYER POLICIES

Section 5.2 uses a deep trunk because Xie et al. (2025) report that clipping fails as policy depth grows. We reproduce that experiment with two hidden layers of width 64, the network of the original PPO study, holding every other setting of Appendix F fixed and running four seeds per environment. Figure 9 gives the learning curves and Table 6 the final returns.

Table 3: Hyperparameters for the Procgen experiments. Rows spanning both columns are shared by every method. The α^{SNR} update rule constants shown are the tuned values at $\epsilon = 0.5$ together with an α ceiling of 2.

	PPO	ALDIPPO (ours)
Environment		
Benchmark	Procgen (Cobbe et al., 2020)	
Distribution mode	easy	
Training levels	200	
Evaluation levels	full distribution	
Environment steps	25×10^6	
Parallel environments	64	
Evaluation	256 episodes every 60 updates	
Optimisation		
Rollout length	256	
Batch / minibatch	16384 / 2048 (8 minibatches)	
Epochs per update	3	
Optimiser	Adam	
Learning rate	5×10^{-4}, linearly annealed to 0	
Discount γ	0.999	
GAE λ	0.95	
Value coefficient c_1	0.5 (value loss clipped)	
Entropy coefficient c_2	0.01	
Gradient-norm clip	0.5	
Advantage normalisation	on (selectors read unnormalised advantages)	
Network		
Encoder	IMPALA-CNN, channels $(16, 32, 32)$, 256 hidden	
Actor / critic	shared trunk, separate linear heads	
Initialisation	orthogonal (policy and value heads)	
Trust region		
Trust-region width ϵ	0.2 (main), 0.5 (robustness)	
Ratio floor	—	$r \geq 10^{-6}$
α range $[\alpha_{\min}, \alpha_{\max}]$	—	$[1, 4]$
Newton step bound	—	unbounded
α EMA β	—	0.9 ($\alpha^{\mathcal{B}2}$, $\alpha^{\mathcal{B}3}$, α^{SNR})
α^{SNR} gate (gain, shift, EV scale)	—	$(40, 0.03, 10)$

Table 4: Procgen held-out test return at $\epsilon = 0.2$ on all sixteen games, reported as mean with standard deviation in parentheses over three seeds at 25M steps. Bold marks the best method in each game. The final row is the mean normalised score, using the easy-mode constants from Cobbe et al. (2020). Mean normalised training scores over the same runs are 0.550 for PPO, 0.567 for SPO, 0.566 for α^{SNR}, 0.597 for $\alpha^{\mathcal{B}1}$, 0.587 for $\alpha^{\mathcal{B}2}$ and 0.595 for $\alpha^{\mathcal{B}3}$.

Environment	PPO	SPO	α^{SNR}	$\alpha^{\mathcal{B}1}$	$\alpha^{\mathcal{B}2}$	$\alpha^{\mathcal{B}3}$
BigFish	1.8 (.9)	2.9 (1.4)	3.2 (1.4)	3.5 (1.3)	**3.9 (2.0)**	3.6 (1.2)
BossFight	7.2 (.5)	7.9 (.3)	**8.1 (.5)**	7.6 (.2)	8.0 (.8)	7.5 (.3)
CaveFlyer	4.0 (.2)	4.8 (.9)	4.5 (.7)	4.8 (.9)	**5.2 (.4)**	4.5 (.7)
Chaser	4.8 (2.5)	4.3 (.4)	5.0 (.5)	5.6 (.7)	5.4 (.5)	**6.5 (.3)**
Climber	3.4 (.0)	4.3 (.3)	4.4 (.3)	4.4 (.3)	**4.6 (.3)**	4.4 (.4)
CoinRun	**8.7 (.2)**	8.6 (.4)	8.3 (.3)	8.4 (.2)	8.5 (.1)	8.3 (.2)
Dodgeball	1.1 (.2)	1.7 (.2)	2.1 (.2)	**2.4 (.3)**	1.8 (.3)	1.9 (.3)
FruitBot	23.4 (.5)	25.2 (.4)	24.2 (.6)	**26.3 (1.0)**	24.1 (.3)	24.7 (.3)
Heist	2.2 (.6)	2.5 (.7)	**2.6 (.8)**	2.3 (1.1)	2.2 (.5)	2.6 (.4)
Jumper	5.7 (.4)	5.8 (.1)	**6.1 (.3)**	5.7 (.4)	6.0 (.0)	6.0 (.4)
Leaper	4.0 (1.4)	2.9 (1.2)	3.8 (1.4)	**5.2 (3.0)**	3.1 (1.3)	3.1 (1.1)
Maze	5.0 (.2)	5.0 (.1)	**5.1 (.1)**	5.0 (.3)	4.9 (.3)	4.9 (.4)
Miner	6.6 (.3)	7.7 (.4)	7.9 (.5)	**8.1 (.2)**	7.9 (.2)	7.9 (.4)
Ninja	5.7 (.3)	5.6 (.1)	5.4 (.2)	6.0 (.3)	**6.2 (.3)**	6.1 (.3)
Plunder	5.9 (.9)	6.7 (.5)	6.5 (.6)	6.7 (1.0)	6.8 (2.0)	**7.6 (.5)**
StarPilot	**35.2 (2.8)**	32.4 (1.6)	33.8 (.8)	33.4 (2.8)	29.7 (2.1)	33.3 (1.8)
Mean normalised	.264 (.007)	.281 (<.001)	.292 (.010)	**.313 (.024)**	.294 (.021)	.297 (.007)

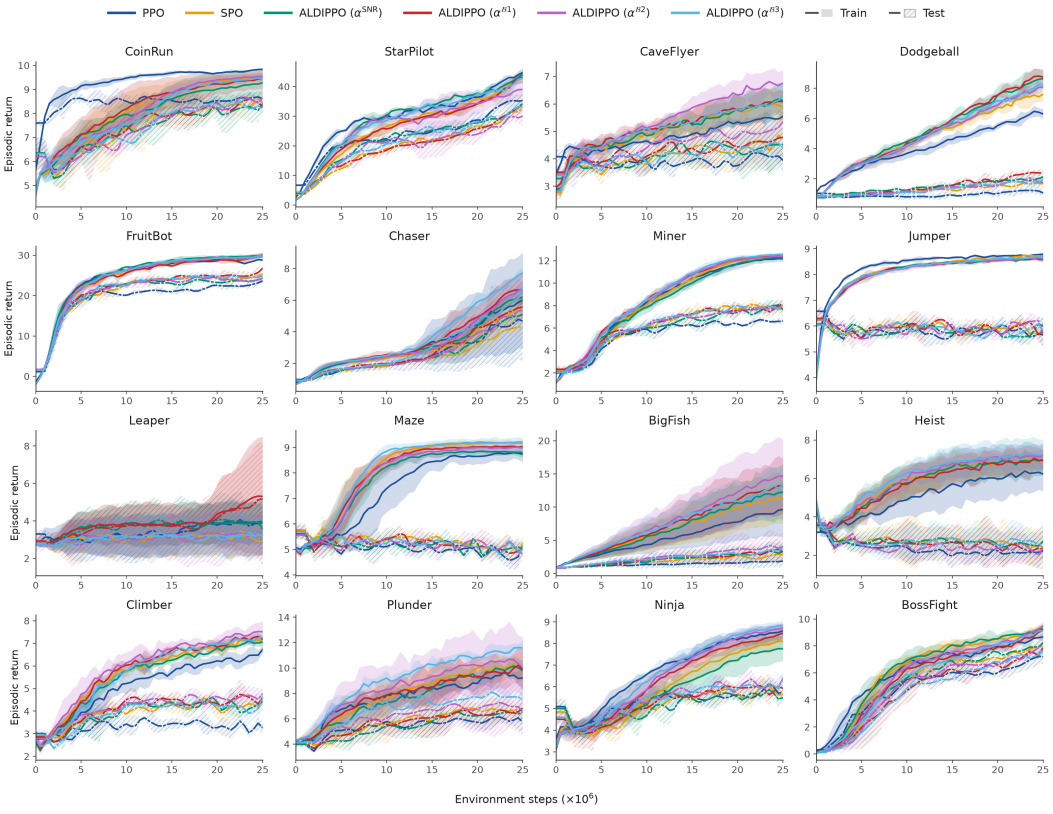

Figure 4: Procgen return at $\epsilon = 0.2$ on all sixteen games, as the mean over three seeds with one standard deviation bands. Solid lines with plain bands are the 200 training levels; dash-dot lines with hatched bands are held-out levels. Final values for every game are in Table 4.

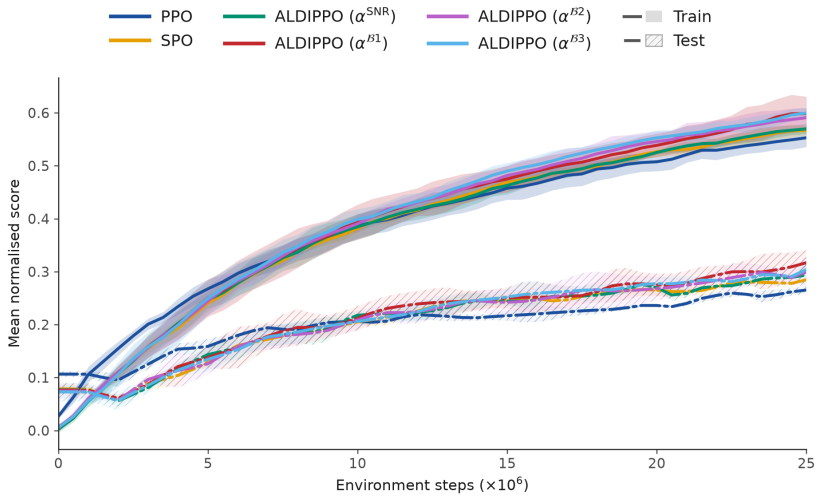

Figure 5: Mean normalised score at $\epsilon = 0.2$ over the sixteen games. The generalisation gap is wide for every method, and ALDIPPO separates from PPO on held-out levels while the training curves stay close. Per-game values are in Table 4.

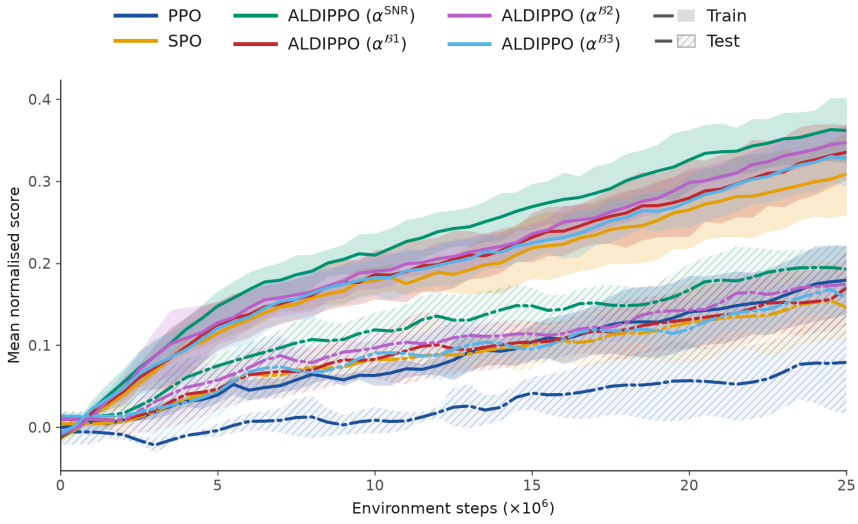

Figure 6: Mean normalised score at $\epsilon = 0.5$ over the eight Procgen games, as the mean over three seeds with one standard deviation bands. Solid lines with plain bands are the 200 training levels; dash-dot lines with hatched bands are held-out levels. PPO separates from every ALDIPPO update rule on held-out levels, while the update rules stay close to one another. α^{SNR} uses the tuned update rule of Appendix E. Per-game curves are in Figure 7.

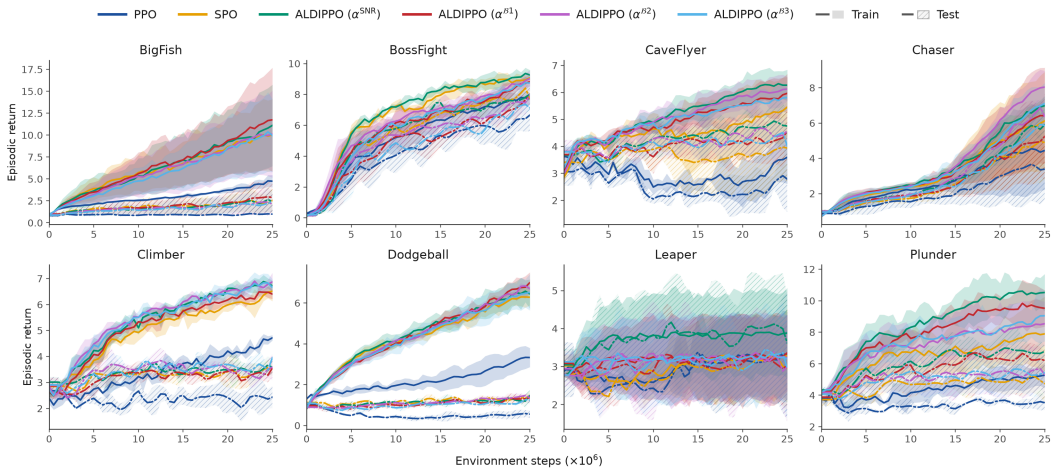

Figure 7: Procgen return at $\epsilon = 0.5$ on the eight games of Table 2, as the mean over three seeds with one standard deviation bands. Solid lines with plain bands are training levels; dash-dot lines with hatched bands are held-out levels. α^{SNR} uses the tuned update rule of this appendix.

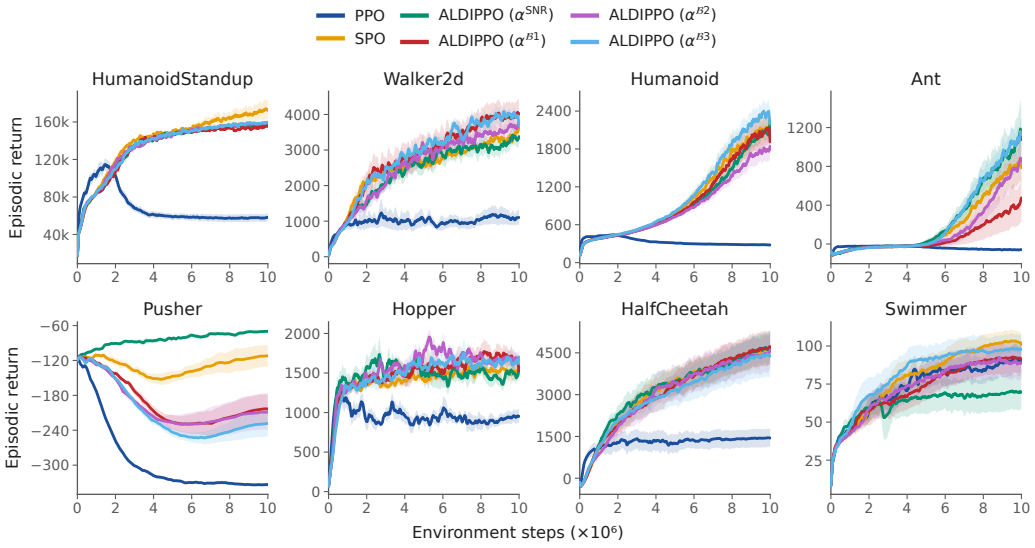

Figure 8: MuJoCo learning curves on all eight environments over six seeds, shown as the mean with a band of one standard error. The first four panels are those of Figure 3. PPO plateaus early or collapses on seven of the eight environments, while the ALDIPPO update rules stay close to one another. The HumanoidStandup panel shows PPO ahead before two million steps and below every other method afterwards.

Table 5: Hyperparameters for MuJoCo Gym. Entries spanning both columns are shared.

	PPO	ALDIPPO (ours)
Environment and budget		
Suite	Gymnasium MuJoCo, version 5	
Environment steps	10^7	
Seeds	6 (1 to 6)	
Action clipping	to the action space	
Observation normalisation	running mean and variance	
Reward normalisation	running, discounted, then clipped to $[-10, 10]$	
Network		
Actor and critic	separate MLPs	
Number of hidden layers	{2,7}	
Hidden dimension	{64,256}	
Activation	$\tanh$	
Initialisation	orthogonal, output gain 0.01 (actor) and 1.0 (critic)	
Policy	Gaussian, state-independent $\log \sigma$ initialised at 0	
Optimisation		
Optimiser	Adam, $\varepsilon_{\mathrm{Adam}} = 10^{-5}$	
Learning rate	5×10^{-4}, and 2×10^{-4} on Pusher and Swimmer	
Learning-rate schedule	linear decay to zero	
Parallel environments	8	
Rollout length	256	
Batch size	2048	
Minibatch size	512	
Update epochs	10	
Discount γ	0.99	
GAE λ	0.95	
Value coefficient c_1	0.5	
Entropy coefficient c_2	5×10^{-3}	
Gradient-norm clip	0.5	
Advantage normalisation	per minibatch	
Value-loss clipping	enabled	
Trust region and update rule		
ϵ	0.2	0.2
α range	—	$[1, 4]$
α update rule	—	α^{SNR}, $\alpha^{\mathcal{B}1}$, $\alpha^{\mathcal{B}2}$, $\alpha^{\mathcal{B}3}$
$\alpha^{\mathcal{B}3}$ smoothing β	—	0.9

Table 6: MuJoCo return with two hidden layers of width 64, reported as mean with standard deviation in parentheses over four seeds at 10M steps. Bold marks the best method in each environment. The final row scales each environment so the worst method is zero and the best is one and reports the median over the eight environments. †HumanoidStandup is reported in thousands.

Environment	PPO	SPO	α^{SNR}	$\alpha^{\mathcal{B}1}$	$\alpha^{\mathcal{B}2}$	$\alpha^{\mathcal{B}3}$
HumanoidStandup†	55.8 (3.2)	151 (7)	181 (31)	**182 (44)**	152 (3)	158 (11)
Walker2d	**3369 (518)**	2779 (182)	2901 (315)	3040 (499)	3020 (66)	2792 (147)
Humanoid	284 (16)	958 (164)	1104 (267)	950 (176)	**1108 (338)**	1044 (357)
Ant	1094 (434)	3033 (184)	**3218 (64)**	2934 (153)	3126 (1007)	3134 (500)
Pusher	−56.6 (4.2)	**−50.4 (13.1)**	−51.0 (5.1)	−50.7 (2.5)	−56.8 (8.1)	−55.1 (4.5)
Hopper	1139 (100)	1654 (91)	1604 (69)	1625 (135)	1571 (58)	**1716 (68)**
HalfCheetah	**4589 (1786)**	3096 (1348)	3217 (1099)	3450 (1757)	1909 (141)	1888 (101)
Swimmer	44.9 (4.9)	45.1 (9.0)	31.2 (6.3)	45.5 (10.2)	**54.0 (12.3)**	50.6 (11.9)
Median normalised	0.018	0.788	**0.860**	0.826	0.758	0.830

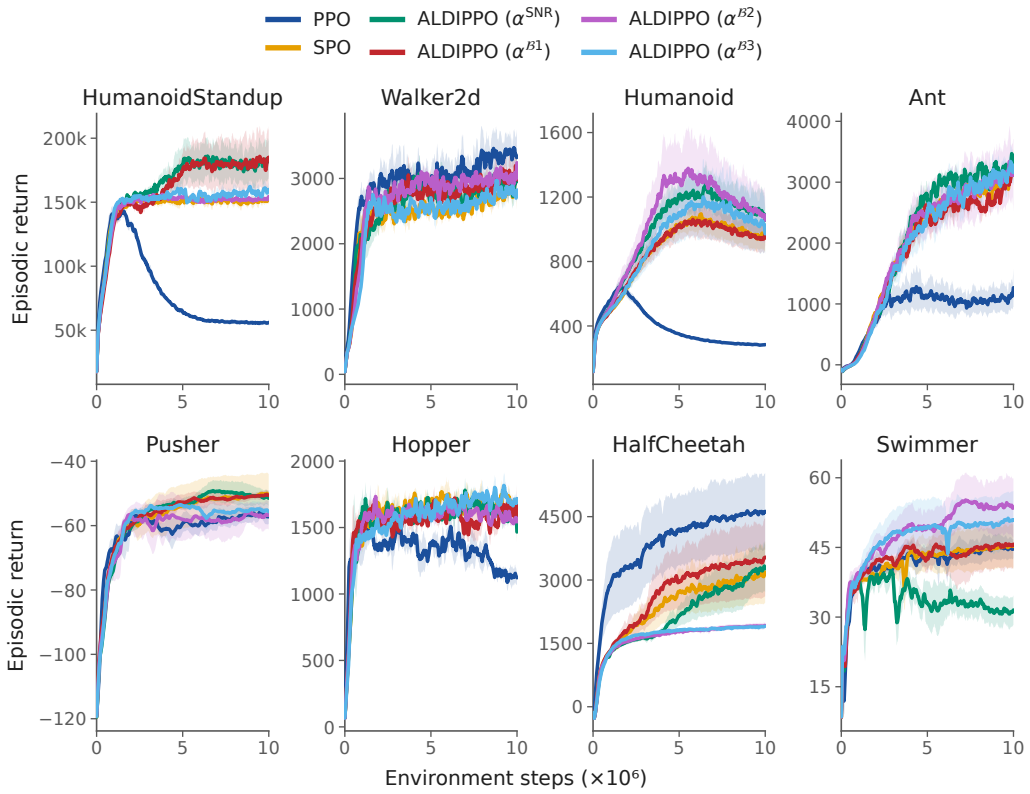

Figure 9: MuJoCo learning curves with two hidden layers of width 64, shown as the across-seed mean with a band of one standard error over four seeds.